Listening. Learning. Leading.®

Research Report
ETS RR–14-33

# Developing Content Knowledge for Teaching Assessments for the Measures of Effective Teaching Study

Geoffrey Phelps

Barbara Weren

Andrew Croft

Drew Gitomer

December 2014

# ETS Research Report Series



Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

# Developing Content Knowledge for Teaching Assessments for the Measures of Effective Teaching Study

Geoffrey Phelps,[1] Barbara Weren,[1] Andrew Croft,[1] & Drew Gitomer[2]

1 Educational Testing Service, Princeton, NJ

2 Rutgers, The State University of New Jersey, Piscataway Township, NJ

This report documents the development of assessments of content knowledge for teaching (CKT) as part of the Measures of Effective Teaching (MET) study, funded by the Bill and Melinda Gates Foundation. The MET study was designed to develop a set of measures that together serve as an accurate indicator of teaching effectiveness. The study was implemented during the 2009–2010 and 2010–2011 school years with more than 3,000 teachers in 6 predominately urban school districts. A total of 5 assessments of CKT were developed, piloted, and then administered as part of the MET study. The CKT assessments focused on the content knowledge used in recognizing, understanding, and responding to the content problems that teachers encounter as they teach a subject. In English language arts (ELA), 2 assessments were developed: 1 for teachers of Grades 4–6 and 1 for Grades 7–9. In mathematics, 3 assessments were developed: 1 for teachers of Grades 4–5, 1 for Grades 6–8, and 1 algebra I. A total of 2,080 final assessments were administered to 1,718 teachers in the 6 participating MET study districts. Assessment results for 194 teachers were excluded due to evidence that assessments were either completed together by 2 or more participants or that insufficient time was devoted to represent a good faith effort at answering the assessment questions. The final sample included 1,886 assessments. Assessment scores included both selected-response and constructed-response (CR) questions. We used information from item level statistics, including percent correct and biserial correlations, to systematically remove poorly performing items in order to improve assessment reliabilities. Item level statistics for each assessment are presented. Descriptive statistics and histograms indicate that participants are well distributed over the range of possible score responses. Assessments had moderate to strong levels of reliability, ranging from 0.69 to 0.83.

**Keywords** Teacher content knowledge; assessment; teaching quality; English language arts; mathematics

doi:10.1002/ets2.12031

The content knowledge for teaching (CKT) assessments described in this report were developed as part of the larger Measures of Effective Teaching (MET) study. The goal of this development effort was to create a set of assessments that measure the types of content knowledge used in teaching practice and then to examine how these CKT assessments are related to other measures of teaching effectiveness. To provide context for the assessment development, the report starts with a brief overview of the larger MET study and of the general theory of CKT that guided both the development of the assessment framework and assessment forms. The second section of the report describes the process of developing an assessment framework, assessment items, and the piloting and revision steps that led to the construction of final assessments (Appendices A–E). The third section of the report presents results from the administration of the assessments to teachers participating in the MET study and includes information on item performance, assessment scoring and reliability, and the general characteristics of participant scores. Appendix F contains item level statistics (i.e., percent correct and biserial correlations) for all items administered.

## Background

### Measures of Effective Teaching Study

The MET study was designed to develop a set of measures that together serve as an indicator of a teacher's impact on student achievement. The study was implemented during the 2009–2010 and 2010–2011 school years with approximately 3,000 teachers in six predominately urban school districts. To limit the need for extensive additional testing of students, the study focused on grades and subjects in which most states currently test students—Grades 4 through 8 mathematics

*Corresponding author*: G. Phelps, E-mail: gphelps@ets.org

and English language arts (ELA) — plus three additional courses that serve as gateways for high school students — Grade 9 algebra I, English, and biology.

The research study was led by more than a dozen academic institutions, not-for-profit institutions, and other educational consultants. Data about the teachers' practice and their students' achievement were collected from a number of sources, including the following:

- Measures of student achievement gains on existing state assessments and supplemental assessments designed to assess higher-order conceptual understanding.
- Classroom observations, scored using multiple observation protocols, and teacher reflections on their practice.
- Measures of teachers' CKT (mathematics and ELA only).
- Surveys of student perceptions of the classroom instructional environment.
- Surveys of teachers' perceptions of the working conditions and instructional support at their schools.

The measures of teachers' CKT in mathematics and ELA were developed during the first year of the study and were administered to teachers participating in the study during the second year.

## Content Knowledge for Teaching

Content knowledge is a long-established basic prerequisite for teaching a subject, and it is an essential requirement for teacher certification (Hill, 2007). However, scholars have argued that teachers need to develop forms of content knowledge that go beyond basic content proficiency to be effective in the classroom. The idea that teachers need to understand and use content in ways particular to teaching was captured in arguments developed by Shulman, who conceptualized *pedagogical content knowledge* (PCK) as comprising

> the most useful forms of representation of those ideas, the most powerful analogies, illustrations, examples, explanations, and demonstrations — in a word, the ways of representing and formulating the subject that make it comprehensible to others. . . . Pedagogical content knowledge also includes an understanding of what makes the learning of specific topics easy or difficult: the conceptions and preconceptions that students of different ages and backgrounds bring with them to the learning of those most frequently taught topics and lessons. (Shulman, 1986, p. 9)

The notion of PCK has been influential to the education field. These ideas have contributed to research that has focused on the central role of content in teaching and that has provided a way to conceptualize teaching as professional work with its own unique knowledge base (Ball, Thames, & Phelps, 2008, p. 392).

Ball et al. (2008) proposed a domain argument that includes the full range of content knowledge used in teaching a subject, with knowledge falling under two major categories: content knowledge and PCK. PCK includes three types of content knowledge that are a mix of content knowledge and knowledge of students, teaching, or curriculum. *Knowledge of content and students* includes, for example, knowledge of particular content misconceptions or confusions that students are likely to exhibit. *Knowledge of content and teaching* includes knowledge of the productive ways of representing particular content to meet student learning needs. *Knowledge of content and curriculum* includes knowledge such as how to use existing curricula to teach specific content.

Content knowledge for teaching also includes three categories of content knowledge that do not involve knowledge of students, teaching, or curriculum. *Common content knowledge* is defined by its shared use in common across adult pursuits other than just teaching. For example, engineers and mathematics teachers both use knowledge of how to solve algebraic problems in conducting their day-to-day work. *Horizon content knowledge* is defined by knowledge of how different content ideas are connected across the content domain, often with more elementary or basic ideas connected to more complex or advanced ideas. For example, one would not expect a high school geometry course to include non-Euclidean geometry, but it is helpful for a teacher teaching the triangle sum theorem to know its connection to non-Euclidean geometry, as well as the mathematical importance of that connection. *Specialized content knowledge* includes types of content knowledge that are only used in teaching. For example, while someone skilled in mathematics may use a particular mathematical strategy or method to solve a set of problems, the teacher must understand a range of strategies that could be used to represent the math in different ways to address different types of student learning needs, differences in the mathematics involved in each of the strategies, and whether such strategies are mathematically valid and generalizable.

These are examples of a type of pure mathematics that are only needed for and used in the work of teaching mathematics and that do not require the additional knowledge about students or content representations that comprise PCK.

Emerging theories about what constitutes content knowledge for teaching have received interest from teacher educators, policy makers, and teachers themselves, because these ideas draw a direct connection between the work of teaching a subject and the content knowledge needed in order to carry out this work. While clearly teachers need to understand the content that their students are learning, evidence that teaching this content can require forms of content knowledge that go far beyond more conventional content proficiency strengthens arguments that teaching is professional work with its own unique professional knowledge base. These ideas also point to the potential and need for assessments that capture forms of content knowledge used in teaching that differ from what is measured on conventional knowledge tests.

## Developing a Design Framework and Content Knowledge for Teaching Assessments

### A Design Framework for Content Knowledge for Teaching Assessments

Building on the literature, we define CKT as the content knowledge used in recognizing, understanding, and responding to the content practices that teachers engage in as they teach a subject. While much of the literature on content knowledge for teaching has focused on distinctions in the types of knowledge used in teaching, we place the emphasis instead on teaching itself and seek to identify the main ways that content knowledge is used in teaching practice (Ball & Bass, 2003). This focus on teaching practice is an important component of current theories of CKT. Domains such as PCK and its subcomponents are defined by the ways in which content is encountered and worked within teaching. In the classroom, PCK exists because teachers work with students' content confusions or with content representations in teaching content. It is the work of teaching and how content is encountered in teaching that distinguishes domains of CKT from content as it is used outside of teaching practice. However, in our work to develop measures of CKT, we have shifted the emphasis in both conceptualization and actual question design from distinctions that differentiate knowledge type to distinctions that differentiate how knowledge is used in teaching. Instead of seeking to develop and design items that measure one type of CKT rather than another, we focus on the practices or tasks of teaching that make up the work of teaching subjects such as ELA and mathematics and organize our domains around those tasks. As a consequence, the assessment questions organize around the work of teaching in whatever mix of the various components of CKT that are needed to address the content problems that teachers need to recognize, understand, or respond to in a particular task.

Our perspective in assessing CKT is defined by the recurrent practices that make up the work of teaching subjects such as mathematics and ELA. These recurrent practices, known as *tasks of teaching*, define what knowledge is needed for teaching and, thus, form a design framework for CKT assessments that directly links content knowledge to teaching practice. One could label this a practice-based approach, because it takes instructional practice as the basis for both organizing knowledge and for designing assessment questions that can measure the content knowledge *used in* teaching.[1]

The CKT assessment design framework focuses on the content practices that teachers encounter in the range of settings and roles that define their work. Most obviously, this involves teachers' interactions with students around the content and learning activities that occur within the classroom. However, teachers also engage in many content practices outside of the classroom as they consider or plan how to teach a topic, make sense of student work products to decide what to do next, share teaching with colleagues, design lessons or curriculum materials, interpret standards, prepare students for state tests, and much more. A practice-based framework for assessing CKT must sample from the full range of content practices that define that work of teaching a subject.

These core practices, or *tasks of teaching*, are the organizing foundation for the CKT assessment design framework. Many tasks of teaching are common across subject areas and grade levels and thus, create a common structure that in turn provides a basis for elaborating the tasks of teaching encountered in a specific subject area and grade level. These elaborated descriptions of tasks of teaching provide the basis for defining the content problems that teachers need to recognize, understand, and respond to and that are the target of particular CKT assessment questions.

The design framework developed for the MET study is outlined in Table 1 and describes the content-specific tasks that comprise the work of teaching at three levels: first a general level that is shared across subjects; next, a level of tasks that are specific to defined content domain; and finally, a level that is defined by integration of the actual content topics that are taught to students with the work of teaching that content. At the highest level, represented in the two left-most columns, the content tasks of teaching are general. Tasks of teaching at this level include, for example,

**Table 1** Content Knowledge for Teaching (CKT) Assessment Design Framework Tasks of Teaching

| Task of teaching | Description | Mathematics | English language arts (ELA) |
|---|---|---|---|
| 1. Anticipating student challenges, misconceptions, partial misconceptions, alternate conceptions, strengths, interests, capabilities, and background knowledge | *This task of teaching describes the work of integrating knowledge about content to be learned and about students, and anticipating how students are likely to interact with the content and content practices. This work is key in planning instruction, both in advance and in the moment as adjustments are made. These anticipations can become the basis for selecting appropriate explanations, examples, or tasks as instructional strategies.* | • Anticipating student challenges in reasoning about and doing mathematics due to the interplay of content demands and students' understanding<br>• Anticipating the impact of limited English language proficiency on students' comprehension of math concepts and on their ability to present mathematical ideas, make mathematical arguments, and give explanations<br>• Anticipating likely misconceptions, partial conceptions, and alternate conceptions about particular mathematics content and practices<br>• Anticipating student interest and motivation around particular mathematics content and practices | • Anticipating student challenges in reading, writing, speaking, or listening due to the interplay of content demands and students' understanding<br>• Anticipating the impact of limited English language proficiency on students' comprehension of text and speech and on their written and spoken expression<br>• Anticipating likely misconceptions, partial conceptions, and alternate conceptions about particular ELA concepts, texts, resources, and processes<br>• Anticipating student interest and motivation around particular ELA concepts, texts, resources, and processes<br>• Anticipating how students' background knowledge, life experiences, and cultural background can interact with new ELA concepts, texts, resources, and processes |
| 2. Evaluating student ideas evident in work, talk, actions, and interactions | *This task of teaching describes the work of making sense of things that students do, say, and produce. It can involve deciding if an idea is valid and/or if an explanation is adequate in a particular context, identifying evidence of partial or alternate understandings, or analyzing errors. Evaluating includes characterizing, analyzing, describing, and judging student work.* | • Evaluating student work, talk, or actions in order to identify conceptions in mathematics, including incorrect or partial conceptions<br>• Evaluating student explanations or arguments for use of appropriate mathematical practices<br>• Evaluating nonstandard responses for evidence of mathematical understanding and in terms of efficiency, validity, and generalizability<br>• Evaluating student representations for evidence of mathematical understanding<br>• Evaluating or comparing multiple solutions to a mathematics problem or explanations of a mathematical concept or procedure<br>• Evaluating discussion among groups of students for evidence of understanding mathematics concepts and practices | • Evaluating student work, talk, or actions in order to identify conceptions, including incorrect or partial conceptions, about ELA concepts, texts, and processes<br>• Evaluating student work, talk, and actions for evidence of strengths and weaknesses in reading, writing, speaking, and listening<br>• Comparing multiple responses to an ELA task<br>• Evaluating discussion among groups of students for evidence of understanding ELA concepts, texts, and processes |

**Table 1** Contiuned

| Task of teaching | Description | Mathematics | English language arts (ELA) |
|---|---|---|---|
| 3. Explaining concepts, procedures, representations, models, examples, definitions, and hypotheses | *This task of teaching describes the crafting and using of appropriate explanations at any point in time when explanations are called for, including in setting purposes for instruction, planning, giving feedback, or in responding to students in the moment. Explaining also includes activities that show, such as modeling.* | • Explaining mathematical concepts, or why a mathematical idea is true<br>• Explaining mathematical procedures<br>• Explaining mathematical representations<br>• Explaining mathematical models<br>• Explaining mathematical examples<br>• Explaining mathematical definitions<br>• Explaining mathematical hypotheses<br>• Explaining why a particular definition, model, or representation serves a particular purpose<br>• Interpreting a particular representation in multiple ways to further understanding<br>• Explaining why a practice, process, or procedure is efficient or appropriate<br>• Explaining the context of a mathematical conjecture and why it is a conjecture | • Explaining literary or language concepts, using definitions, examples, and analogies when appropriate<br>• Explaining processes of reading, including why certain processes are appropriate for particular texts and/or tasks<br>• Explaining processes of writing, including why certain processes are appropriate for particular tasks<br>• Explaining processes of listening and speaking, including why certain processes are appropriate for particular topics, groups, and/or task<br>• Explaining the use of representations to support understanding or development of text<br>• Explaining hypotheses about texts |
| 4. Creating and adapting resources for instruction (examples, models, representations, explanations, definitions, hypotheses, procedures) | *This task of teaching describes the work of inventing new instructional tools or adapting available ones to fit particular purposes. While curricular resources can be a source of material for adaptation, student-generated materials can also be adapted into opportunities for learning. The creation or adaptation of instructional tools is thought of as having an instructional goal, examples of which are in the subject-specific bulleted lists. Examples are selected for a reason—to introduce new material, to showcase certain attributes of content, or to challenge students' current conceptions.* | • Creating and adapting examples to introduce a concept or illustrate an idea or to demonstrate a strategy, procedure, or practice<br>• Creating and adapting examples that support particular mathematical strategies or to address particular student questions, misconceptions, or challenges with content<br>• Creating and adapting representations or models to support students' mathematical understanding<br>• Creating and adapting multiple representations or models to support students' mathematical understanding across representations<br>• Creating and adapting representations or models that support multiple interpretations<br>• Creating and adapting definitions to fit instructional purposes<br>• Adapting student-generated conjectures to support instructional purposes<br>• Creating and adapting procedures for working with content | • Creating and adapting examples or model texts to introduce a concept or to demonstrate a literary technique or a reading, writing, or speaking strategy<br>• Creating and adapting examples or model texts to develop understanding of a concept, literary technique, or literacy strategy, or to address particular student questions, misconceptions, or challenges<br>• Creating and adapting representations (e.g., graphic organizers, think-alouds) to support understanding or development of text or other products<br>• Creating and adapting definitions of ELA concepts and processes to fit instructional purposes<br>• Adapting student contributions to fit instructional purposes<br>• Creating and adapting procedures for writing, speaking, listening, or working with texts<br>• Creating and adapting analogies to support student understanding of ELA concepts, texts, and processes |

**Table 1** Contiuned

| Task of teaching | Description | Mathematics | English language arts (ELA) |
|---|---|---|---|
| 5. Evaluating and selecting resources for instruction (examples, models, representations, explanations, definitions, hypotheses, procedures) | *This task of teaching describes the work of evaluating instructional resources and selecting those most appropriate for a particular instructional purpose. This work can occur through advance planning or during instruction. Teachers may evaluate and select from curricular materials or from potential resources generated during the course of instruction, including work generated by students.* | • Evaluating and selecting examples to introduce a concept or illustrate an idea or to demonstrate a strategy, procedure, or practice<br>• Evaluating and selecting examples that support particular mathematical strategies or to address particular student questions, misconceptions, or challenges with content<br>• Evaluating and selecting representations or models to support students' mathematical understanding, or in terms of validity, generalizability, or fit to the concept, calculation, etc. to be represented<br>• Evaluating and selecting representations or models that support multiple interpretations<br>• Evaluating and selecting definitions to fit instructional purposes<br>• Evaluating and selecting explanations of mathematical concepts for potential to support mathematical learning or in terms of validity, generalizability, or explanatory power<br>• Evaluating and selecting procedures for working with mathematics content<br>• Evaluating and selecting student-generated conjectures to support instructional purposes<br>• Evaluating and selecting resources for their potential to support mathematical learning | • Evaluating and selecting examples or model texts to introduce a concept or to demonstrate a literary technique or literacy strategy<br>• Evaluating and selecting examples to develop understanding of a concept, literary technique, or literacy strategy, or to address particular student questions, misconceptions, or challenges with content<br>• Evaluating and selecting representations (e.g., graphic organizers, think-alouds) to support understanding or development of text or other products or in terms of fit to the concept to be represented<br>• Evaluating and selecting definitions of ELA concepts and processes to fit instructional purposes<br>• Evaluating and selecting explanations of ELA concepts, texts, and processes for their potential to support learning<br>• Evaluating and selecting procedures for writing or working with text<br>• Evaluating and selecting resources to support particular instructional goals<br>• Evaluating and selecting analogies to support student understanding of ELA concepts, texts, and processes |
| 6. Developing questions, activities, tasks, and problems to elicit student thinking | *This task of teaching includes the work of developing and using questions, tasks, and problems to engage students in developing content understanding. The development of questions, tasks, and problems is used to make student understanding visible and to build deeper and more accurate understanding.* | • Creating or adapting problems or questions with the potential to elicit student mathematical thinking, discussions, justifications, or explanations<br>• Creating or adapting problems that support particular mathematical strategies and practices<br>• Creating or adapting questions, activities, tasks, or problems that demonstrate desired mathematical characteristics<br>• Creating or adapting classes of problems that address the same mathematical concept or that systematically vary in difficulty and complexity<br>• Creating or adapting questions, activities, or tasks to elicit evidence that students have a particular mathematical understanding or skill | • Creating or adapting questions with the potential to elicit discussion or analysis of a text or a literary concept<br>• Creating or adapting prompts or questions with the potential to elicit productive student writing<br>• Creating or adapting activities or tasks that support the development of a particular literary understanding or skill<br>• Creating or adapting questions, activities, or tasks to elicit evidence that students have a particular literary understanding or skill |

**Table 1** Contiuned

| Task of teaching | Description | Mathematics | English language arts (ELA) |
|---|---|---|---|
| 7. Evaluating and selecting student tasks (questions, tasks, problems) to elicit student thinking | *This task of teaching includes the work of evaluating and selecting things for students to work on or interact with. The evaluation and selection of student tasks is thought of as having an implicit instructional goal and relies on anticipations about how students will understand and interact with the tasks.* | • Evaluating and selecting problems or questions to elicit student mathematical thinking, justifications, or explanations<br>• Evaluating and selecting problems that support particular mathematical strategies and practices<br>• Evaluating and selecting questions, tasks, or problems that demonstrate desired characteristics<br>• Evaluating and selecting classes of problems that address the same mathematical concepts or that systematically vary in difficulty or complexity<br>• Evaluating and selecting questions, activities, or tasks to elicit evidence that students have a particular mathematical understanding or skill | • Evaluating and selecting questions, activities, or tasks to elicit discussion about a specific text or literary concept<br>• Evaluating and selecting questions to elicit productive student writing<br>• Evaluating and selecting questions, activities, or tasks to support the development of a particular literary understanding or skill<br>• Evaluating and selecting questions, activities, or tasks to elicit evidence that students have a particular literary understanding or skill |
| 8. Doing the work of the student curriculum | *This task of teaching includes doing the student level tasks that make up the intended curriculum. Doing the student level work is not uniquely demanded by teaching, but is an essential part of the content work necessary to do in the course of teaching, especially in preparation for assigning student tasks. This work is often incorporated into or prerequisite for doing other tasks, such as anticipating student difficulties or selecting problems or tasks.* | • Doing the work that will be demanded of the students as part of the intended curriculum | • Doing the work that will be demanded of the students as part of the intended curriculum |

anticipating student difficulties, evaluating student work, and explaining concepts. At this level, tasks of teaching are applicable to all subjects regardless of topic or grade level and, thus, provide a general organization that can guide development for any assessment of CKT. Tasks of teaching that are specific to a subject are indicated in the final two columns.

The third level of the hierarchy specifies the content topic level distinctions that are the focus of each of the tasks of teaching. For example, a task of teaching, such as anticipating student challenges, misconceptions, partial misconceptions, alternate conceptions, strengths, interests, capabilities, and background knowledge (see Table 1), might focus on anticipating likely student misconceptions around number concepts, rational numbers, proportional reasoning, algebra, and so forth. The third level of the framework represents the particular ways that teachers work with the topical content that they are responsible for teaching. We have not attempted to represent this third level of the framework in Table 1 because of the level of detail that would be involved. In an ideal form of test development, all of the ways that content intersects with tasks of teaching would be identified and there would be a clear mapping between this elaborated map of content teaching and the test items that sample from this domain. This level of domain development was not undertaken for the MET assessments. However, the actual items, which are presented in the appendices, provide a useful illustration of how tasks of teaching involve work with particular topical content and the range of topical content that was addressed at the third level of the framework.

The CKT assessments, and in particular the third level of the CKT framework, are guided by attention to general topic level strands that define the content that is taught and learned at the grade levels that are assessed. As a practical matter, identifying these general topic strands can present challenges in developing a valid and relevant assessment. Often states and even districts have quite different standards for the student content that is taught, the grade level at which this content is taught, and to what level of content proficiency. These differences make it difficult to justify the specific content that should be emphasized in a CKT assessment. It is worth noting that recent changes in student standards toward a more commonly shared set of student standards (e.g., the Common Core State Standards [CCSS]) could have a major impact on the development of CKT assessments of mathematics and ELA. If, for example, CCSS were fully adopted by all schools and states in the United States, then the CCSS would become the basis for the third level of the hierarchy. At this point, however, no such shared basis exists, and the third level of the hierarchy is by necessity a set of decisions and compromises in emphasizing content that is widely shared enough to be deemed a necessary component of CKT for a particular subject and grade span.

The list of tasks of teaching (see Table 1) is not taken to be comprehensive, although we have made efforts to generalize across content domains. This list is not intended to imply that this is all teaching consists of, but it is a result of our focus on the *content* problems teachers must contend with in the work of teaching. The tasks on the list are not intended to be seen as entirely exclusive of one another. In many cases, the work of teaching might include multiple tasks, or a single task might potentially be described in more than one way. For example, a piece of student work completed in class might first be evaluated by the teacher, then become an example that is evaluated and selected because it highlights a particular point that the teacher has in mind due to anticipating certain misconceptions about the content. Larger tasks of teaching not represented here may be made up of smaller tasks of teaching. Responding to students, for example, is a combination of evaluating what the student has said and done, choosing an appropriate response, and then delivering that response. The tasks of teaching listed here would include evaluation of the student work and creation of an explanation, example, model, representation, or something else as appropriate to the student understandings or misunderstandings at hand.

Tasks of teaching are the organizing basis for the design framework because each task represents a natural unit of teaching practice that has been identified in research or other relevant analyses of teaching practice as being a critical part of teaching. As natural units of practice, they often represent the coordination of different knowledge domains of CKT. For example, a task of teaching could simultaneously draw on common content knowledge, specialized content knowledge, and knowledge of content and teaching. Rather than artificially segmenting instructional problems to align with a discrete knowledge domain of CKT, the tasks of teaching attempt to capture how teachers actually work with content in their day-to-day and moment-to-moment practice. This supports validity claims that assessments built on this framework measure the intended knowledge and reasoning and also increase the face validity of assessment tasks by situating them in contexts that are perceived by teachers as authentic and relevant.

**Table 2** Advisory Group Members

| Consultant | Affiliation | Content area |
|---|---|---|
| Margarita Calderón | Johns Hopkins University | English language arts |
| Pam Grossman | Stanford University | English language arts |
| Carol Lee | Northwestern University | English language arts |
| James Marshall | University of Georgia | English language arts |
| Deborah McCutchen | University of Washington | English language arts |
| P. David Pearson | University of California, Berkeley | English language arts |
| Deborah Ball | University of Michigan | Mathematics |
| William Bush | University of Louisville | Mathematics |
| Heather Hill | Harvard University | Mathematics |
| Erik Jacobson | University of Georgia | Mathematics |
| Rebecca McGraw | University of Arizona | Mathematics |
| Judit Moschkovich | University of California, Santa Cruz | Mathematics |
| Denise Spangler | University of Georgia | Mathematics |
| Jon Star | Harvard University | Mathematics |

## Assessment Development

### *Development Teams*

The practice-based approach that led to the definition of the CKT construct was also the basis for developing assessments of CKT. For each subject area, a development team was created to generate items that were directly linked to the framework and, thus, to teaching that subject. Each development team comprised a mix of individuals, including professional assessment developers at Educational Testing Service (ETS), researchers working on the CKT project, teacher educators, and educators working directly with K-12 students. The goal of item development was to create assessment tasks that elicited the knowledge and reasoning used to recognize, understand, and act on content problems encountered in these tasks of teaching.

Advisory groups were recruited to provide consultation and advice during the framework and item development. Members of the advisory committees (Table 2) were nationally recognized scholars in the field of mathematics or ELA teacher education. These consultants provided feedback and guidance during the early development of the CKT task of teaching framework. They were also given the opportunity to review and provide feedback on all of the items that were piloted in 2010.

The development of the teacher knowledge measures was a joint effort of researchers from the University of Michigan (UM) and researchers and assessment specialists from ETS and was built directly on the extensive experience gained from the Learning Mathematics for Teaching (LMT) project at the UM. The proposed plan was to develop a series of six assessments of teacher knowledge for teachers of the following subjects and grades:

- Mathematics Grades 4 – 5
- Mathematics Grades 6 – 8
- Algebra I
- ELA Grades 4 – 5
- ELA Grades 6 – 8
- English Grade 9

However, a series of discussions among the research team and advisors led us to decide to limit development to two ELA assessments, as the natural break was one in which the respective curricula emphasized skill development versus literature. Therefore, two ELA teacher knowledge measures were developed: one for teachers of ELA in Grades 4 – 6 and the second for teachers of ELA in the Grades 7 – 9. Three mathematics assessments were developed, as described above. During the project proposal phase, consideration was also given to developing a measure of content knowledge for teaching high school biology, but after discussions with the MET leadership, this was not deemed to be feasible given the constraints of the project budget, the short development timeline, and limited expertise in this area.

The assessments of teacher knowledge were designed to be delivered over the Internet, allowing teachers to take assessments when they chose. From the beginning, we recognized the challenge of finding a reasonable balance between

Ms. Hupman is teaching an introductory lesson on exponents. She wants to give her students a quick problem at the end of class to check their proficiency in evaluating simple exponential expressions. Of the following expressions, which would be least useful in assessing student proficiency in evaluating simple exponential expressions?

A) $3^3$
B) $2^3$
C) $2^2$
D) All of these are equally useful in assessing student proficiency in evaluating simple exponential expressions.

**Figure 1** Sample mathematics content knowledge for teaching (CKT) item.

Ms. Rice begins a unit on memoir writing by reading a passage from a literary model. She then asks students to complete a warm-up activity to help them generate ideas for their own writing.

For each assignment, indicate whether or not it will help students focus their brainstorming on generating a memoir.

| | **Will help focus brainstorming** | **Will not help focus brainstorming** |
|---|---|---|
| (A) Write a poem about the ways you have changed, using the form "I used to be…but now I am…" | | |
| (B) Write a sequence of sentences describing some of your experiences, beginning each sentence with the phrase "I remember." | | |
| (C) Write a few adjectives that describe your personality. | | |
| (D) Write down some of your favorite foods and describe what you like about them. | | |

**Figure 2** Sample English language arts (ELA) content knowledge for teaching (CKT) table item.

assessment length, participant burden, and adequate assessment reliabilities. Given that there were no stakes associated with assessments in the context of the MET study and participation was voluntary, the assessment length was largely driven by a decision to keep the assessments relatively short to reduce participant burden. If these teacher knowledge assessments had been intended to be used to discriminate among teachers for the purpose of high stakes decisions such as tenure and firing, then very high levels of assessment reliability and associated precision for individual classifications would be necessary. Achieving high levels of reliability is typically accomplished by creating long assessments. However, reliabilities can also be enhanced by making use of performance data from pilot administrations to select high performing items and strike a balance in design that maximizes assessment reliabilities and minimizes participant burden. For these particular assessments, our goal was to create assessments that had reliabilities above 0.70 and could be completed in an hour or less.

### *Question Types*

The question development work built directly on the experiences of the LMT project at the UM. The LMT measures comprise two types of questions: single-selection multiple-choice, in which the single best answer is chosen from among the four answer choices given, and what we refer to as *table questions*, in which a common stimulus is presented with two or more answer choices (presented in columns), and a separate response must be given for each scenario presented in the rows of the table. Typically, table items present a yes or no response option. Likewise, the majority of the questions included in the MET assessments of teacher knowledge used these same two response types. Sample MET questions are shown in Figures 1 (example of a single-selection multiple-choice question) and 2 (example of a table question).

For the purposes of clarity, in all subsequent descriptions of the assessments and in the discussion of the data analyses, each response to a selected-response question will be referred to as an item. Thus, as can be seen in Figures 1 and 2, a single-selection multiple-choice question corresponds to a single item, but a table question corresponds to several items, one for each row in the table (e.g., the sample table question corresponds to four items). Each table question in the MET teacher knowledge assessments corresponds to three to six items. (A discussion about the independence of the items within a table question is included in the scoring results section below.)

In addition to the two types of selected-response questions described previously, four of the five measures of teacher knowledge (i.e., all except algebra I) also included a small number of open-ended or constructed-response (CR) questions. One compelling reason to include CR questions in assessments like these is that they can capture evidence of certain types of knowledge (e.g., generative reasoning) that is not available from selected-response questions. Although many large-scale assessments (teacher and student) have not included CR tasks because of the significant costs associated with scoring the tasks with human raters, recent developments in natural language processing have made it possible to conceive of scoring short constructed responses using automated scoring technologies. Thus, the proposal for the development of teacher knowledge measures included small-scale efforts to (a) develop CR questions that reliably measure some aspect of the CKT construct and (b) investigate the feasibility of using automated scoring technologies to score the responses. (Since the algebra I assessment was only administered to teachers at one grade level, Grade 9, the expected sample size was too small to support building automated scoring models, and thus, no CR items were included in the algebra I assessment.)

### *Item-Writing Procedures*

This assessment development was not a standard effort in two important ways. First, for this study, the assessment development teams were simultaneously working within and iteratively developing a theoretical framework for the assessments. In more traditional assessment development, the assessment design and specifications emerge from a consensus-building process among content-expert development committee members. These specifications often take the form of lists of important content topics, rather than an overarching framework, and are defined prior to actual assessment development work. Second, because the content domains for basic mathematics knowledge and reading ability are fairly well specified, development of tests of teacher knowledge have traditionally followed a rather straightforward procedure, creating questions that focus on content topics that include, and to some degree encompass, the student curriculum for the grade span. This requires assessment developers to have knowledge of their subject area and expertise in constructing assessments but not necessarily deep understanding of teaching. However, the starting point for development of the MET CKT assessments was teaching practice itself, as defined by the tasks of teaching. Therefore, the assessment development required simultaneous attention to teaching practice, teacher content knowledge as it is used in teaching practice, and more standard conceptions of the content knowledge represented by the student curriculum. Not only does this complicate the work because additional, deeper knowledge is required, but it also requires conventionally trained assessment developers to learn how to attend to dimensions of teaching in assessment development.

We found that to do this work, assessment development teams needed a solid grounding in instruction in their subject matter area, and both needed to develop the skills to examine particular tasks of teaching and to identify the knowledge entailed by those tasks. Consequently, the mathematics and ELA assessment development teams each comprised researchers whose work focused on the study of teaching, subject matter experts, teacher educators, and assessment specialists, many who also had teaching experience. The development teams also collaborated with external groups with specific expertise around teaching, such as a group of professional writers who both work directly with students and provide professional development to teachers of creative writing.

Both of the development teams met regularly, virtually, and face to face, throughout the first year of the study to develop a pool of questions that could be used in piloting and as examples to generate additional questions. Members of the development teams drafted questions that focused on the tasks of teaching and core content topics for each particular assessment. All questions were reviewed and discussed in detail multiple times by the development teams to ensure that the questions would elicit the desired evidence of teacher knowledge and to inform the continuous development of the CKT framework. Questions that were deemed acceptable by the development teams were then reviewed for compliance with editorial standards and best practices for assessment development. All revisions to the questions were documented and tracked to enable us to study the development process at a future time.

### *Pilot*

A pool of more than 200 questions was developed across the five measures. During the summer of 2010, two forms of each of the five assessments were assembled and piloted with a convenience sample of practicing teachers, teaching the

**Table 3** Summary of Question Types, Item Exclusion, and Assessment Reliability

| Assessment form | Total SR items | SR items excluded | SR item reliability | Total CR items | Final reliability |
|---|---|---|---|---|---|
| Mathematics 4–5 | 38 | 0 | 0.74 | 2 | 0.76 |
| Mathematics 6–8 | 46 | 0 | 0.82 | 2 | 0.83 |
| Algebra I | 37 | 2 | 0.77 | 0 | 0.77 |
| ELA 4–6 | 51 | 5 | 0.71 | 2 | 0.74 |
| ELA 7–9 | 48 | 3 | 0.66 | 2 | 0.69 |

*Note.* SR = selected response; CR = constructed response; ELA = English language arts.

appropriate subject area and grade level. Extensive efforts were made to recruit potential participants for the pilot administration of the MET CKT items. More than 13,000 teachers were sent invitations to participate. Potential participants were recruited in numerous ways, including direct individual contact using information provided by supporting school districts; charter school operators; and relevant organization LISTSERVs, newsletters, and Web sites. Approximately 1,600 teachers consented to participate in the pilot study, with a total of 1,751 assessments completed. Those teaching at the elementary level were given the option to take both mathematics and ELA assessments.

Participants in the MET study were not allowed to take the pilot assessments. The pilot assessments, along with a short demographic survey, were administered using the same Web-based assessment delivery platform that was used for the MET study administrations. Each assessment was designed to take approximately 1 hour to complete and was administered to 150–200 participants. Detailed analyses were then conducted of the item performance to examine item difficulty, discrimination, and latency.

In parallel with the piloting, a small number of cognitive interviews were conducted with teachers (mostly from New Jersey), using a subset of the questions. These interviews were designed to elicit teachers' thinking as they worked through the questions and to provide information about potential sources of confusion and/or construct-irrelevant variance. Interviews were used to show whether and how effectively the intended constructs were actually being measured by the assessment questions. The results of these interviews were used to support the revision and refinement of the CKT questions included in the assessment forms for the study administration.

Also in parallel with the pilot administration, all items were provided to an external advisory committee for review and comment.

Data from the psychometric analyses and cognitive interviews, along with comments from the external reviewers, were used by the development teams to select and revise the questions for inclusion in the assessments administered to the MET study participants. The assessments as administered are summarized in Table 3, with the actual questions administered on each form presented in Appendices A–E.

### *Final Assessments*

Assessment reliability (Cronbach's raw alpha) ranged from 0.69 to 0.83 for the CKT assessments based on selected-response and CR items (Table 3). Certain selected-response items were excluded from scoring and item analysis through a process elaborated in the Item Analysis and Assessment Reliability section of this report.

A number of questions appeared on multiple assessments. Eight questions appeared on both ELA assessments. Six questions were common to the mathematics 4–5 and mathematics 6–8 assessments. Four questions appeared on both the mathematics 6–8 and algebra I assessments. One question was common to the mathematics 4–5 and algebra I assessments. As discussed in the results, these common items provided an opportunity to investigate differences in item performance across assessments and groups of teachers.

## Administration and Sample

The administration of MET assessments of teacher knowledge occurred during the 2010–2011 school year. We administered 2,080 CKT assessments to teachers among six districts participating in the MET study. Teachers received unique URLs to access and complete assessments online at any time during the administration period. The assessment for teachers of mathematics in Grades 4 and 5 was administered in fall of 2010, and the other four assessments were administered in early 2011. Assessments were administered to teachers at the MET-specified grade level (i.e., Grade 4 and Grade 5

**Table 4** Exclusion Codes by Assessment

| | Assessment | | | | | |
|---|---|---|---|---|---|---|
| | Math 4–5 $n = 465$ *n* (%) | Math 6–8 $n = 398$ *n* (%) | Algebra I $n = 148$ *n* (%) | ELA 4–6 $n = 635$ *n* (%) | ELA 7–9 $n = 434$ *n* (%) | Total $n = 2{,}080$ *n* (%) |
| Exclusion category | | | | | | |
| Missing responses | 0 (0) | 3 (1) | 0 | 2 (<1) | 4 (1) | 9 (<1) |
| Latency/response pattern | 8 (2) | 5 (1) | 3 (2) | 13 (2) | 4 (1) | 33 (2) |
| Unusual agreement | 60 (13) | 17 (4) | 2 (1) | 65 (10) | 8 (2) | 152 (7) |
| Final scored sample | 397 (85) | 373 (94) | 143 (97) | 555 (87) | 418 (96) | 1,886 (91) |

*Note.* ELA = English language arts. Percentages shown are with respect to the assessment total *n*.

mathematics teachers received the mathematics 4–5 assessment). The algebra I assessment was administered only to Grade 9 algebra I teachers.

## Excluding Cases From the Scored Sample

We reviewed the assessment results in order to identify those that could not be used to generate reliable scores (e.g., cases of incomplete assessments in which responses were given only to the first few questions or cases in which the entire assessment was completed within a few minutes). In addition, as part of the review process, we noticed patterns in the timing data that suggested possible problems, including collaboration among groups (either pairs or larger groups) of test takers. We then conducted a systematic and more detailed analysis of assessment data. As a result, we excluded 194 assessment results from the final sample. These cases were not included in item analyses and were not individually scored. Cases were excluded for a variety of reasons, but the largest number of exclusions were cases of unusual agreement, instances in which two or more teachers appearing to have collaborated in ways that rendered it impossible to attribute the scores to a particular individual. We excluded additional cases for other reasons that created concern about good faith effort, including questionable response patterns and response times that were not sufficiently long. The excluded cases are summarized in Table 4.

The following descriptions address each of the rules that we applied for this set of analyses. While the overall exclusion rate is approximately 9%, exclusions were more frequent for elementary level assessments.

### *Missing Responses*

In these cases, the teacher submitted fewer than 75% of required item responses to complete the assessment (i.e., 25% or more of that teacher's responses were missing). In nearly all cases, these were sequential responses at the end of the assessment, as teachers were required to enter a response in order to proceed. Thus, these individuals exited the assessment prior to completion.

### *Latency and Response Pattern*

Teachers completed the assessment at a rate deemed to be too fast to have considered items and thought about response choices. Assessments for which the total latency was less than 5 minutes or for which either half of the assessment was completed in less than 2 minutes and 30 seconds were excluded. Time spent on CR items was not included in this criterion.

We checked each teacher's set of responses for predominant submission of any single response option that suggested that items were not given serious consideration. If over 80% of overall responses or the entire second half of an assessment were selected as a single response option (e.g., the second choice for each item was selected), then the assessment results were flagged. Latency data were used as confirmatory evidence for these types of response patterns. All of these cases were also flagged for speedy assessment completion.

### *Unusual Agreement*

We used statistical analysis to compare all possible pairings of teachers in the entire sample to detect unusual agreement. The methodology for detecting unusual agreement was developed at ETS as a means of test security; the numerical value

generated is known as the K-index (Holland, 1996; Lewis & Thayer, 1998). The software matched all possible pairs of teachers' incorrect response choices on multiple-choice items. The K-index is

> an estimate of the chance that at least the observed number of matching incorrect responses would occur if the two test takers being compared were working independently on the section of the test being analyzed. If this probability is sufficiently small, it is taken as an indicator of "unusual agreement" between the two sets of incorrect responses being compared. (Lewis & Thayer, 1998, p. 1)

Other characteristics of this detection method include:

- Matched correct responses are not included in these analyses because they may indicate that respondents independently know the content and, thus, provide weak evidence of unusual agreement.
- Omitted responses are ambiguous and are not considered evidence for unusual agreement.
- In order to be able to conduct these analyses, there is a minimum number of multiple-choice items for which response data are available (each of the five CKT assessments met this minimum threshold).
- Item difficulty is not controlled for in this analysis.

For this analysis, if the K-index for a pair of assessment results was less than .01, then those results were flagged for unusual agreement. As described below, flagging by itself was not sufficient to exclude a score. Given a large sample size and all possible pairs of teachers for each assessment, many pairs would be identified by chance. Therefore, we collected further evidence to confirm unusual agreement between teachers, identified in pairs, whose results had a K-index below .01.

A. Same district: The teachers in a pair are in the same district, indicating they have the opportunity to complete the assessment together.
B. Same school: The teachers in a pair are in the same school, indicating they have the opportunity to complete the assessment together.
C. End time: The teachers in a pair completed the assessment within 10 minutes of each other.
D. Similar answers on multiple-choice questions: Either teacher in an identified pair must not differ in more than two incorrect responses for multiple-choice questions.
E. Similar answers on table questions: The K-index is determined only using the multiple-choice items. In this analysis, we also checked the consistency of responses on the table questions. Teachers in an identified pairing must not differ in more than two incorrect table item responses. This criterion provided confirmatory evidence of unusual agreement.

In summary, if the K-index is below .01 and criteria A and one of B, C, D, or E are satisfied, we excluded that case on the basis of unusual agreement.

Recognizing that the K-index analysis has weak power when applied to pairs of test takers with few incorrect items, we also reviewed all of the assessment results for teachers in the same schools. We compared their assessment end times, scores, and latencies for the entire assessment and scores and latencies for each half of the assessment. Four pairs of teachers who completed ELA assessments were flagged for completing the assessment at the same time (within 5 minutes of each other) with matching responses and amounts of time spent per item. These teachers were not identified by the K-index because each teacher had only one incorrect response to selected-response items. On the basis of these other data, these were judged to be four cases of unusual agreement, and the eight associated assessment results were excluded from the final analyses.

In sum, 152 assessments were excluded based on the unusual agreement criterion, and the remaining 42 cases due to other criteria used to identify scores that were deemed to provide invalid information on the scored responses.

## Final Data Used for Scoring

The final sample that could be used for scoring contains 1,613 individual teachers who completed 1,886 assessments (273 teachers completed both a mathematics and an ELA assessment). Table 5 summarizes the administration by assessment and by district, respectively.

**Table 5** Final Sample of Content Knowledge for Teaching (CKT) Assessments by District

| | Assessment | | | | | |
|---|---|---|---|---|---|---|
| District | Math 4–5 *n* (%) | Math 6–8 *n* (%) | Algebra I *n* (%) | ELA 4–6 *n* (%) | ELA 7–9 *n* (%) | Total *n* (%) |
| A | 104 (26) | 58 (16) | 18 (13) | 131 (24) | 66 (16) | 377 (20) |
| B | 95 (24) | 83 (22) | 59 (41) | 125 (23) | 119 (28) | 481 (25) |
| C | 85 (21) | 113 (30) | 25 (17) | 141 (25) | 120 (29) | 484 (26) |
| D | 19 (5) | 0[a] | 21 (15) | 13 (2) | 21 (5) | 74 (4) |
| E | 94 (24) | 29 (8) | 20 (14) | 116 (21) | 18 (4) | 277 (15) |
| F | 0[a] | 90 (24) | 0[a] | 29 (5) | 74 (18) | 193 (10) |
| Total | 397 | 373 | 143 | 555 | 418 | 1,886 |

*Note.* ELA = English language arts.

[a]Assessment not administered in this district.

## Measures of Effective Teaching Content Knowledge for Teachers Assessments Results

### Scoring Results

We investigated different scoring models and weightings for some of the item types used in the CKT assessments. In this section, we discuss the considerations and decisions made and then present the final score distributions for each of the assessments.

#### *Selected-Response Items*

Two types of selected-response questions were used in the CKT assessments: multiple-choice questions and table questions. Each multiple-choice question was given equal weight and counted as 1 point because each required a single selected response.

We investigated various models of scoring the multiple-response table questions. We examined intra-item correlations (between the items within each table question) and found, for nearly all table questions, that items within a table question were not highly correlated with each other; that is, each item appeared to measure a discrete aspect of CKT. We concluded that each item (within a table question) should be counted as 1 point, as there was no evidence that item performance violated assumptions of local independence. Had there been evidence supporting a lack of local independence, we would have adopted an alternate weighting strategy so as not to inflate reliability estimates artificially. Each table question counted as 3–6 points, depending on the number of items within the table question.

Before calculating final scores for selected-response items, we analyzed item level data (biserial correlations) and assessment reliability. We systematically removed weakly discriminating selected-response items from each assessment before calculating final scores. We describe the item exclusion process in more detail below.

#### *Constructed-Response Items*

To complete the analysis of the assessments of teacher knowledge, we calculated scores for CR items. For each item, we defined a set of concepts or propositions that represented a correct and complete response to the prompt. We then developed a set of scoring rules that matched the presence of some subsets of concepts with a score on the item. For example, if the correct and complete answer for a problem required the presence of concepts A, B, and C, then scoring rules were created to assign a score when A, B, and C were present (i.e., the highest score possible for the item), as well as scores for subsets of concepts in the response (e.g., A and B or B only). Human scorers had to evaluate whether each of the key concepts was present in the response.

Similarly, the automated scoring engine determined the presence or absence of evidence of each of the scoring concepts by determining the extent to which the test-taker response had the same semantic meaning as any or all of the targeted concepts and then applied the respective scoring rules. A large number of prescored responses were used to build models of acceptable and unacceptable responses, and then the computer system used natural language processing techniques to interpret whether different phrasing had the same semantic meaning as a target concept. Varied degrees of success

**Table 6** Summary of Constructed-Response (CR) Items

| Assessment | Item sequence number | Number of concepts | Initial score range | Quadratic weighted kappa | Mean scale score[b] |
|---|---|---|---|---|---|
| Mathematics 4–5 | 20 | 5 | 0–2 | n/a[a] | 0.77 |
| Mathematics 4–5 | 21 | 2 | 0–2 | 0.93 | 0.99 |
| Mathematics 6–8 | 22 | 5 | 0–4 | 0.87 | 1.88 |
| Mathematics 6–8 | 23 | 6 | 0–2 | 0.77 | 1.51 |
| ELA 4–6 | 18 | 4 | 0–3 | 0.68 | 1.13 |
| ELA 4–6 | 22 | 14 | 0–4 | 0.83 | 1.47 |
| ELA 7–9 | 12 | 4 | 0–2 | 0.87 | 1.31 |
| ELA 7–9 | 18 | 5 | 0–3 | 0.70 | 1.19 |

*Note.* ELA = English language arts.

[a]Item machine scored only, so a measure of interrater reliability (quadratic weighted kappa) cannot be computed.

[b]Initial score range used to generate raw score by two human raters. An average of the two rater scores was then scaled between 0 and 3 points to calculate a final score for constructed-response items.

were achieved with the automated scoring of the CR items, with the general conclusion that it is possible to develop CR items that measure some aspects of CKT and can be reliably scored using automated scoring techniques. Significant additional research and development around these techniques and the applicability to measures of CKT is needed. With one exception, the reported data are based only on human scores (each response scored by two humans). The scoring of one mathematics item was based entirely on numerical criteria (no text or language evaluation) and was machine scored. The interrater reliability of human scorers was calculated using the quadratic weighted kappa. The reliability of the human scoring for all CR items was judged to be adequate for the purposes of this study (Table 6).

We compared two scaling approaches of CR items to determine the degree to which they factor into final total scores. The first approach was to scale CR items at equal value with selected-response items. The second approach was to scale CR items on a 0- to 3-point scale, which is triple the value of selected-response items. We calculated correlations for each approach with total scores including selected-response items only. Correlations were 0.99 or greater using the single-point scaling approach and 0.97 or greater using the triple-point scaling approach. Factoring this result in with the amount of effort required to answer CR items, we decided to use the 0- to 3-point scale to calculate final scores.

A summary of CR item scoring is provided in Table 6.

## Item Analysis and Assessment Reliability

### *Item Level Analyses*

We applied the item level scoring methods for each item type to create total raw scores. For each item, we calculated percent of responses correct (i.e., p-plus value) and biserial correlations as measures of item performance. We investigated the impact on the reliability of each assessment when excluding items with poor discrimination, as measured by item biserial correlations (i.e., the relationship between respondents' performance on one item and performance on the entire assessment). First, items, including table items, with negative biserial correlations were excluded. Second, items were eliminated one at a time, starting with the item with the lowest biserial correlation. If the assessment reliability (measured by Cronbach's raw alpha) increased by 0.01 or greater when an item was excluded, then that item was removed from the subsequent score computations, and the process was repeated. If the alpha did not increase by 0.01 or more from the previous value following the exclusion of an item, then that item was not removed from the subsequent calculations, and the process of excluding items from that assessment was concluded. The number of excluded items by form is reported in Table 3. More detailed information about each item is noted in the tables in Appendix F.

### *Including Constructed-Response Items in Scale Scores*

We investigated the impact on the reliability of each assessment of including CR items, as measured by Cronbach's raw alpha. Assessment reliabilities improved by 0.01 or greater after including CR items. There was no change in the algebra I assessment reliability because it does not contain CR items. No CR items were identified as low-performing or in need of consideration for exclusion.

**Table 7** Range of Item Difficulty

| Assessment | *N* | Minimum % correct | Maximum % correct | *N* < 25% correct | *N* > 75% correct |
|---|---|---|---|---|---|
| Mathematics 4–5 | | | | | |
| Multiple-choice items | 14 | 11.6 | 73.3 | 4 | 0 |
| Table items | 24 | 27.7 | 96.0 | 0 | 8 |
| Mathematics 6–8 | | | | | |
| Multiple-choice items | 14 | 9.4 | 67.6 | 3 | 0 |
| Table items | 32 | 35.4 | 98.1 | 0 | 16 |
| Algebra I | | | | | |
| Multiple-choice items | 19 | 14.0 | 76.9 | 3 | 2 |
| Table items | 18 | 30.1 | 100.0 | 0 | 11 |
| ELA 4–6 | | | | | |
| Multiple-choice items | 17 | 26.9 | 91.9 | 0 | 4 |
| Table items | 34 | 33.3 | 96.0 | 0 | 17 |
| ELA 7–9 | | | | | |
| Multiple-choice items | 16 | 20.3 | 81.6 | 1 | 2 |
| Table items | 32 | 10.5 | 99.3 | 2 | 18 |

*Note.* ELA = English language arts.

**Table 8** Comparison of Percent Correct on Common Subsets of Content Knowledge for Teaching (CKT) Items

| Assessment forms | Items shared across forms | Percentage of items with average scores higher on higher-grade form |
|---|---|---|
| Mathematics 4–5 and Mathematics 6–8 | 9 | 100 |
| Mathematics 6–8 and Algebra I | 12 | 92 |
| ELA 4–6 and ELA 7–9 | 19 | 79 |

*Note.* ELA = English language arts.

The results from the item exclusion analysis and inclusion of constructed response items are summarized in Table 3. The resulting assessment reliabilities range from 0.69 to 0.83.

## Comparing Item Difficulty Across Assessments

As the item level data suggest (Table 7), mathematics assessments were, in general, more difficult than the ELA assessments. No item on the ELA 4–6 assessment was answered correctly by less than 25% of respondents, and more than 40% of items on the ELA assessments were answered correctly by 75% or more of respondents. The mathematics 4–5 assessment contained the highest percentage of items (29%) that were answered correctly by less than 25% of respondents and the lowest percentage (21%) of items for which more than 75% of respondents answered correctly. However, the algebra assessment had proportions similar to the ELA assessments regarding items that were difficult or easy for the majority of participants.

## Comparing Item Difficulty Across Grade Levels

Teachers at different grade levels will have varying experience in teaching particular concepts. It is also true that teachers in middle school tend to have been more likely than elementary teachers to major in the content area that they are teaching (Gitomer, 2007). Thus, we explored whether consistent differences existed in performance between these two groups on the same CKT items shared across tests taken by lower-grade and higher-grade teachers (e.g., mathematics 4–5 vs. mathematics 6–8). We found a consistent pattern where a greater percentage of teachers at the higher-grade levels answered common items correctly compared with teachers at the lower-grade levels. This discrepancy in group performance is greater on common mathematics items than on ELA items (Table 8).

## Examining Score Distributions

Overall, mathematics assessments were more difficult than the ELA assessments, as indicated by the mean percent correct scores reported in Table 9.

**Table 9** Descriptive Statistics of Content Knowledge for Teaching (CKT) Assessment Scale Scores

| Assessment | *n* | Minimum | Maximum | *M* | *SD* |
|---|---|---|---|---|---|
| Mathematics 4–5 | 397 | 20.5 | 93.2 | 52.2 | 14.4 |
| Mathematics 6–8 | 373 | 22.6 | 94.2 | 62.1 | 14.9 |
| Algebra I | 143 | 22.9 | 97.1 | 61.5 | 14.3 |
| ELA 4–6 | 555 | 30.8 | 89.4 | 66.4 | 11.7 |
| ELA 7–9 | 418 | 31.4 | 88.7 | 64.7 | 10.5 |

*Note.* ELA = English language arts.

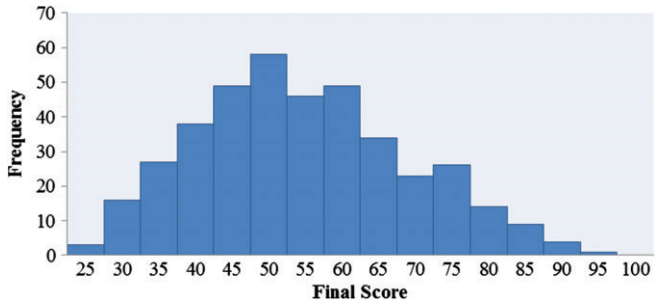



**Figure 3** Mathematics 4–5 score distribution (% correct).

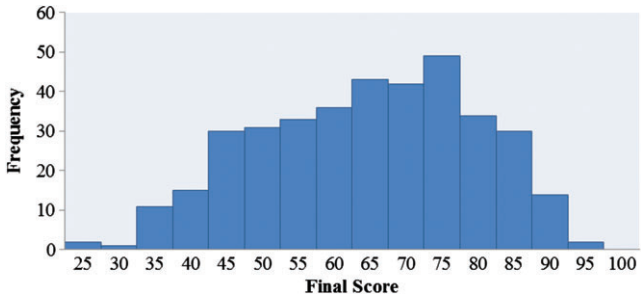



**Figure 4** Mathematics 6–8 score distribution (% correct).

To further examine the distribution of these scores across the assessment samples, we present score distributions on the percent correct scale scores in Figures 3–7. The distribution of scores is near normal for all assessments with a very broad range of scores. The mathematics 4–5 assessment distribution is positively skewed with a large number of low scores, and the ELA distributions are negatively skewed with a large number of high scores. The other mathematics assessments do not appear to be significantly skewed.

## Relationship Between Performances on Both Elementary Level Assessments

We designed the administration of the CKT assessments so that we could make comparisons across assessments. We analyzed the performance of teachers who completed both assessments at the elementary level in ELA and mathematics. After removal of unreliable scores, there remained a total of 271 teachers who had valid scores for both elementary assessments. We found a significant correlation between performances on both assessments ($r = 0.49$, $p < .01$). This relationship is represented graphically in Figure 8. It is also noteworthy that the ELA and mathematics assessments are clearly measuring different domains of knowledge, since the lack of correlation cannot be explained simply by uncorrelated measurement error (the correlation disattenuated for measurement error = 0.68).

Table 10 summarizes the performance on the mathematics and ELA CKT assessments by quartile. While quartile rank is similar for many teachers across assessments, there is a substantial number of teachers who are relatively strong on one test and relatively weak on another. For example, of the 85 teachers in the lowest math quartile, 17 (20%) have ELA scores in the upper two quartiles.

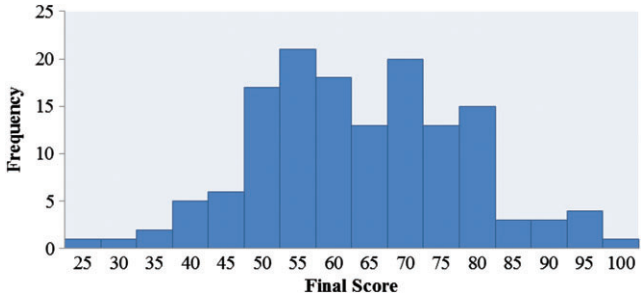



**Figure 5** Algebra I score distribution (% correct).

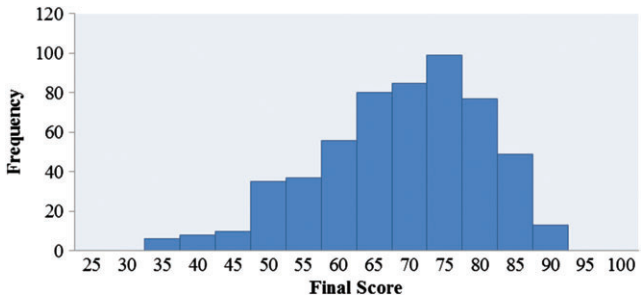



**Figure 6** English language arts (ELA) 4–6 score distribution (% correct).

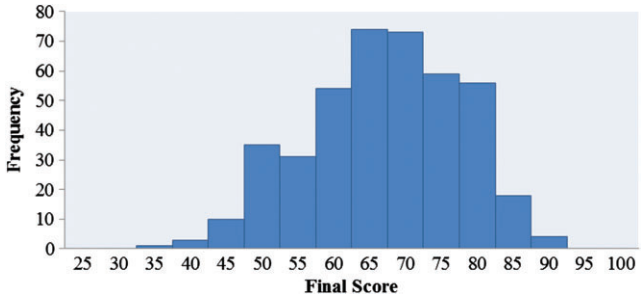



**Figure 7** English language arts (ELA) 7–9 score distribution (% correct).

## Summary

The CKT development conducted as part of the MET study achieved a number of important goals. We highlight three accomplishments and findings: (a) the development of an assessment framework, (b) the generation of reliable assessments suitable for differentiating teachers across a range of CKT proficiency, and (c) practical considerations for how these assessments might be used to assess teacher quality.

Prior to the MET study, the assessment of CKT was carried out as part of research projects, primarily in mathematics, and largely without a well-articulated assessment design framework. One goal of the MET study was to propose a framework that could support identifying domains of CKT both within a given subject and across different subject areas. An important contribution is the CKT framework itself, along with the evidence supported by the development project that the framework can be used to support assessment development. The framework provides a starting point that can guide development efforts in other subjects and grade levels. The high-level categories provide a general organization that allows for making connections across different assessments. The lower-level, subject-specific categories provide examples that can guide the development of analogous or unique subject-specific tasks as needed. The CKT framework holds the promise of being generative for future assessment development.

The CKT framework also focuses attention on teaching. By definition, assessments of CKT are focused on the content knowledge that is needed to recognize, understand, or act on the content problems encountered in teaching. For this reason, CKT questions typically incorporate a teaching scenario or setting that test takers must consider in answering the

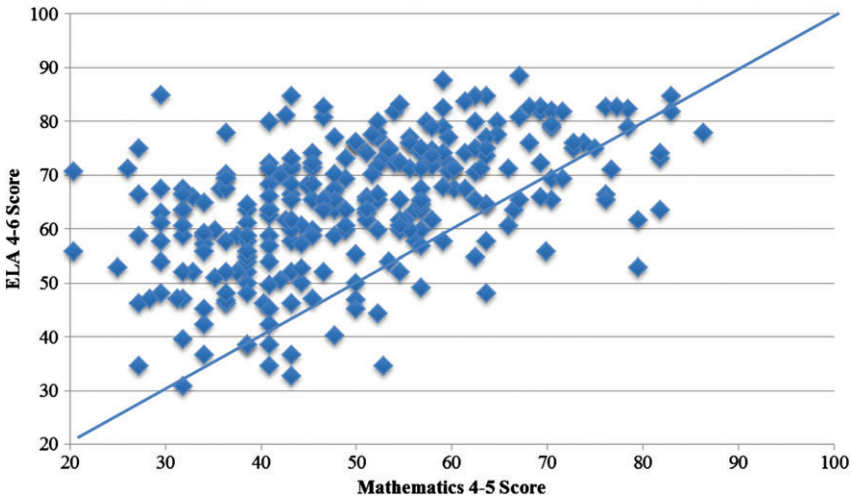



**Figure 8** Scores on elementary level assessment for teachers taking both the mathematics and English language arts (ELA) assessments.

**Table 10** Elementary Assessment Score Quartile Crosstabulation

| | ELA scale score quartiles | | | | |
|---|---|---|---|---|---|
| Math scale score quartiles | Q1 | Q2 | Q3 | Q4 | Total |
| Q1 | 51 | 17 | 14 | 3 | 85 |
| Q2 | 18 | 23 | 16 | 8 | 65 |
| Q3 | 8 | 16 | 25 | 16 | 65 |
| Q4 | 5 | 12 | 14 | 25 | 56 |
| Total | 82 | 68 | 69 | 52 | 271 |

*Note.* ELA = English language arts.

content problem. Organizing the CKT assessment framework around tasks of teaching provides direct guidance for the types of teaching contexts that need to be represented in the assessment tasks. Our success in using the CKT framework to develop assessment questions across five tests in ELA and mathematics demonstrates the utility of a framework designed around tasks of teaching.

Another important contribution of this work is the actual set of assessments supported by the evidence that they can be used to generate reliable test scores that differentiate teachers across a wide range of the score distribution. The reliabilities for these five tests ranged from a low of 0.69 for Grades 7 – 9 ELA to a high of 0.83 for Grades 6 – 8 mathematics. While these reliabilities may seem low, it is important to realize that these assessments are comparable in their measurement quality to existing tests of teacher knowledge. The finding that the CKT tests for ELA were less reliable than for mathematics is also true for traditional content tests for teachers (ETS, 2010, p. 58), likely indicating domain difference rather than an inherent shortcoming of the CKT measures. Further, the CKT assessments are half the length of knowledge tests in current use. If the tests had been of equivalent lengths, the reliabilities would have been comparable to the more traditional knowledge tests, such as those from the *PRAXIS*® series. The CKT tests also provided adequate reliability across the wide range of ability. For all tests, a substantial number of teachers answered less than 50% of the items correctly and a substantial number answered more than 75% of the items correctly. This finding suggests that these tests would provide information on teachers' CKT across a large range of the knowledge continuum.

Finally, the results from the study also provide valuable information on how these assessments function in real school settings. The MET study involved teachers from six districts, a wide range of schools, and many different backgrounds. In a testing situation with no stakes attached, where teacher scores remained anonymous, and in which teachers were allowed to self-administer the assessment, evidence showed that the majority of teachers took these assessments seriously. Most teachers provided a good faith effort, spending the time needed to answer all assessment questions. This effort was true for teachers who received high scores and those who received low scores. This finding is encouraging, as it suggests that these assessments can be used to gather reliable information under conditions that do not include strong incentives for participation.

At the same time, a substantial number of teachers did not complete assessments as directed. On the elementary tests, nearly 15% of teachers appear to have completed the assessment with another teacher. This was the case even though teachers themselves were not being evaluated. It is unlikely that assessments such as these could be self-administered under conditions where teachers were being evaluated or other stakes were attached to their performance. In such cases, the number of collaborating teachers would likely increase along with the incentive to achieve a high score. Testing in situations where scores were public or used for consequential decisions would need to take place in a secure and proctored testing environment.

The results of this development effort are an assessment design framework and a set of assessments that reliably measures the types of content knowledge used in teaching practice. These assessments can be used to examine how measures of CKT are related to other measures of teaching effectiveness. They can be administered and studied in a variety of professional contexts. Future development of these assessments, or the creation of additional ones, can improve how we measure teaching quality and contribute to student learning.

## Acknowledgments

The authors would like to acknowledge the help from many people in the development of the content knowledge for teaching (CKT) assessments. Mark Thames led the initial work developing questions for the mathematics assessments. Barbara Weren was instrumental in all stages of the development, analysis, revisions, and administration of all assessments. Many individuals contributed to writing assessment questions, including Benjamin Baehr, Hyman Bass, Deanna Birdyshaw, Allison Brettschneider, Lindsay Brown, Leah Bullock, Matthew Burgess, Richard Chilcoat, Samantha Caughlan, Susan Dasch, Jenny DeMonte, Arne Jakobsen, Amy Johnson, Nadia Kalman, Minsung Kwon, Sarah Porter, Sarah Scott, and Elida Wylie. We also want to thank the following individuals who assisted with analysis, editing, study recruitment, and study administration: Michaela Arzt, Courtney Bell, Juana Betancourt, John Blackmore, Jason Bonthron, Andrew Croft, Michael Ecker, Jim Fife, Rebecca Gleeson, James Halliday, Barbara Hames, Erik Jacobson, Jyoti Kamal, Steve Meshanko, Anita O'Brien, Rob Rarich, Frank Rijmen, Melanie Schine, Jana Sukkarieh, Barbara Suomi, Amy Swauger, Sailesh Vezzu, and Mike Wagner.

## Notes

1 The distinction between organizing assessment by tasks of teaching rather than types of CKT is rather subtle. For a full discussion of this shift of perspective and what this particular approach offers for assessing content knowledge, see Phelps, Howell, and Kirui (2013).

## References

Ball, D. L., & Bass, H. (2003). Toward a practice-based theory of mathematical knowledge for teaching. In B. Davis & E. Simmt (Eds.), *Proceedings of the 2002 annual meeting of the Canadian Mathematics Education Study Group* (pp. 3–14). Edmonton, AB: CMESG/GDEDM.

Ball, D. L., Thames, M. H., & Phelps, G. (2008). Content knowledge for teaching: What makes it special? *Journal of Teacher Education*, *59*(5), 389–407. doi:10.1177/0022487108324554

Educational Testing Service. (2010). *Praxis technical manual*. Princeton, NJ: Author.

Gitomer, D. H. (2007). *Teacher quality in a changing policy landscape: Improvements in the teacher pool* (Policy Information Report). Princeton, NJ: Educational Testing Service.

Hill, H. C. (2007). Mathematical knowledge of middle school teachers: Implications for the No Child Left Behind policy initiative. *Educational Evaluation and Policy Analysis*, *29*(2), 95–114. doi:10.3102/0162373707301711

Holland, P. W. (1996). *Assessing unusual agreement between the incorrect answers of two examinees using the K-index: Statistical theory and empirical support* (Research Report No. RR-96-07). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1996.tb01685.x

Lewis, C., & Thayer, D. T. (1998). *The power of the K-index (or PMIR) to detect copying* (Research Report No. RR-96-49). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1998.tb01798.x

Phelps, G., Howell, H., & Kirui, D. (2013). *Assessing the content knowledge used in teaching*. Manuscript submitted for publication.

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, *15*(2), 4–14.

# Appendix A

## Mathematics 4–5 Assessment

**Table A1** Item/Question Sequence Number for Mathematics 4–5 Assessment

| Item/question sequence number | Answer key | Item/question sequence number | Answer key |
|---|---|---|---|
| 1 | A | 10d | Provides |
| 2 | C | 10e | Provides |
| 3a | Not problematic | 11 | B |
| 3b | Problematic | 12 | B |
| 3c | Not problematic | 13 | B |
| 3d | Not problematic | 14a | Provides |
| 3e | Problematic | 14b | Does not provide |
| 4 | C | 14c | Provides |
| 5 | A | 14d | Provides |
| 6 | A | 14e | Does not provide |
| 7a | Provides | 15 | B |
| 7b | Does not provide | 16 | B |
| 7c | Provides | 17 | C |
| 7d | Provides | 18 | C |
| 7e | Provides | 19a | Can |
| 8 | A | 19b | Can |
| 9 | B | 19c | Cannot |
| 10a | Does not provide | 19d | Can |
| 10b | Provides | 20 | Open-ended response |
| 10c | Does not provide | 21 | Open-ended response |

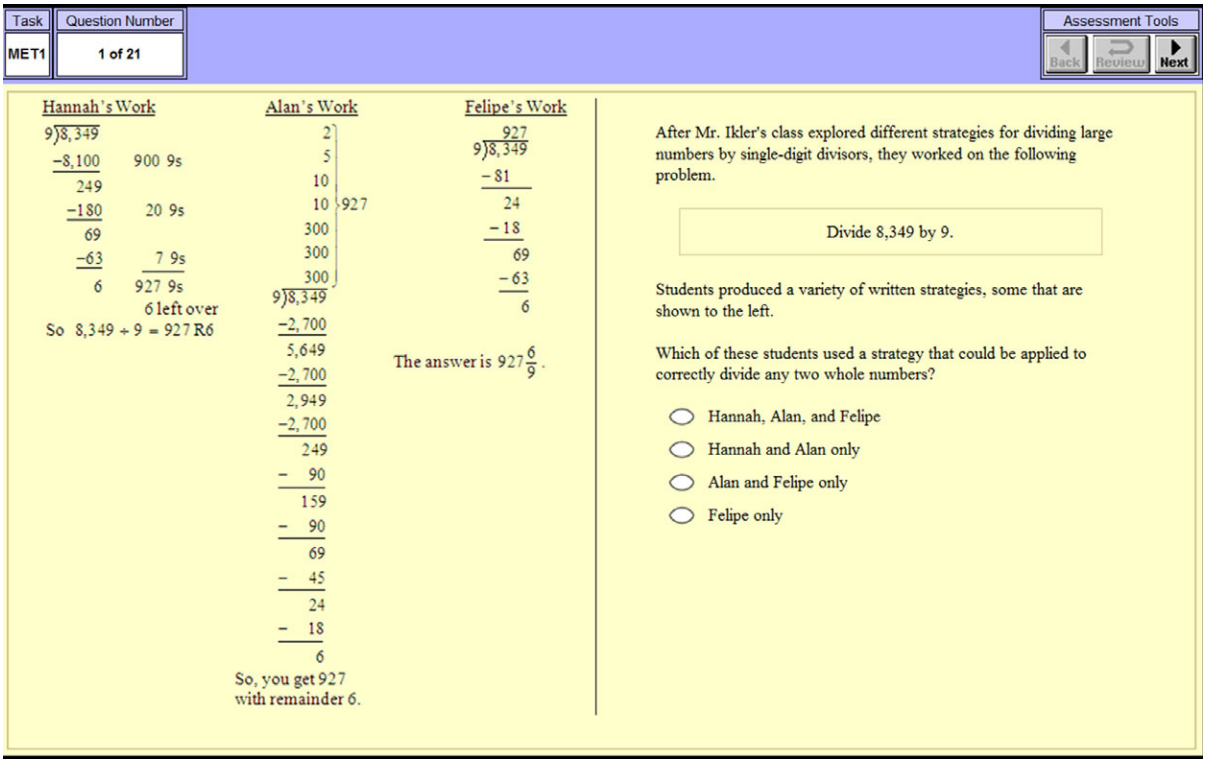



**Figure A1** Question #1.

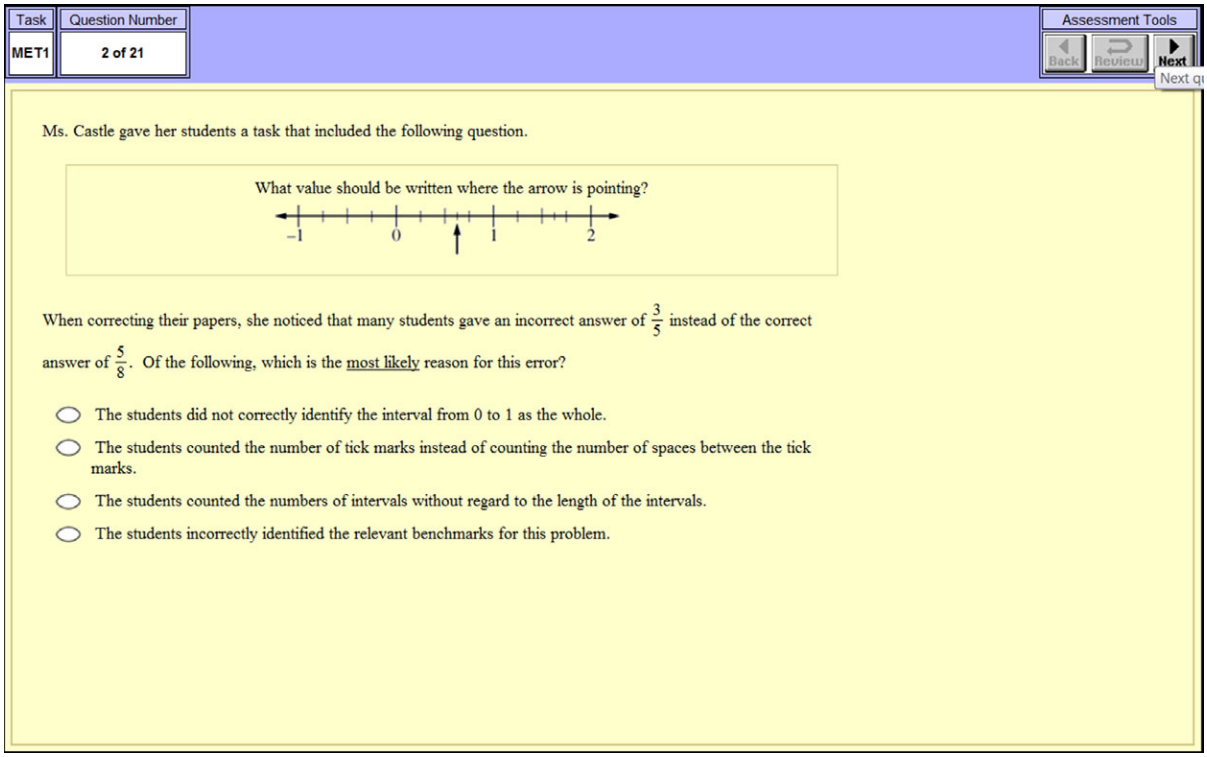



**Figure A2** Question #2.

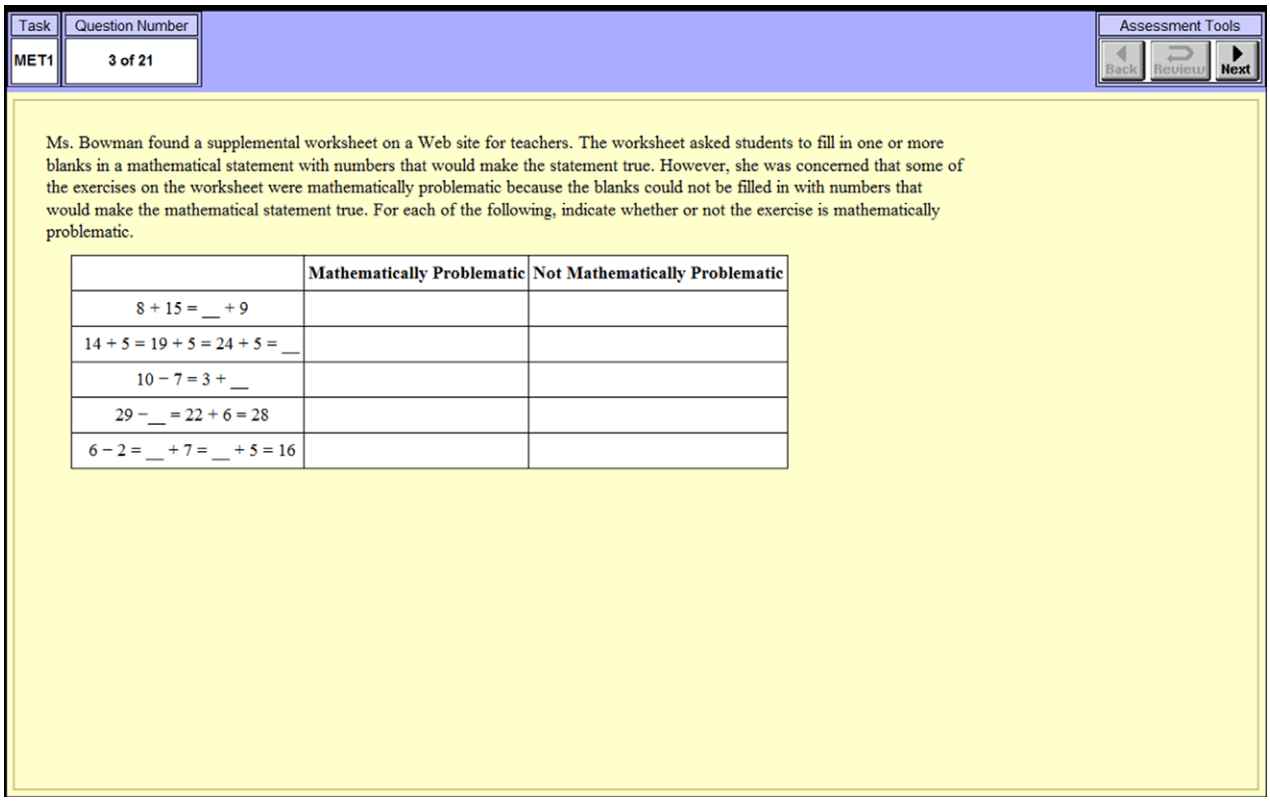



**Figure A3** Question #3.

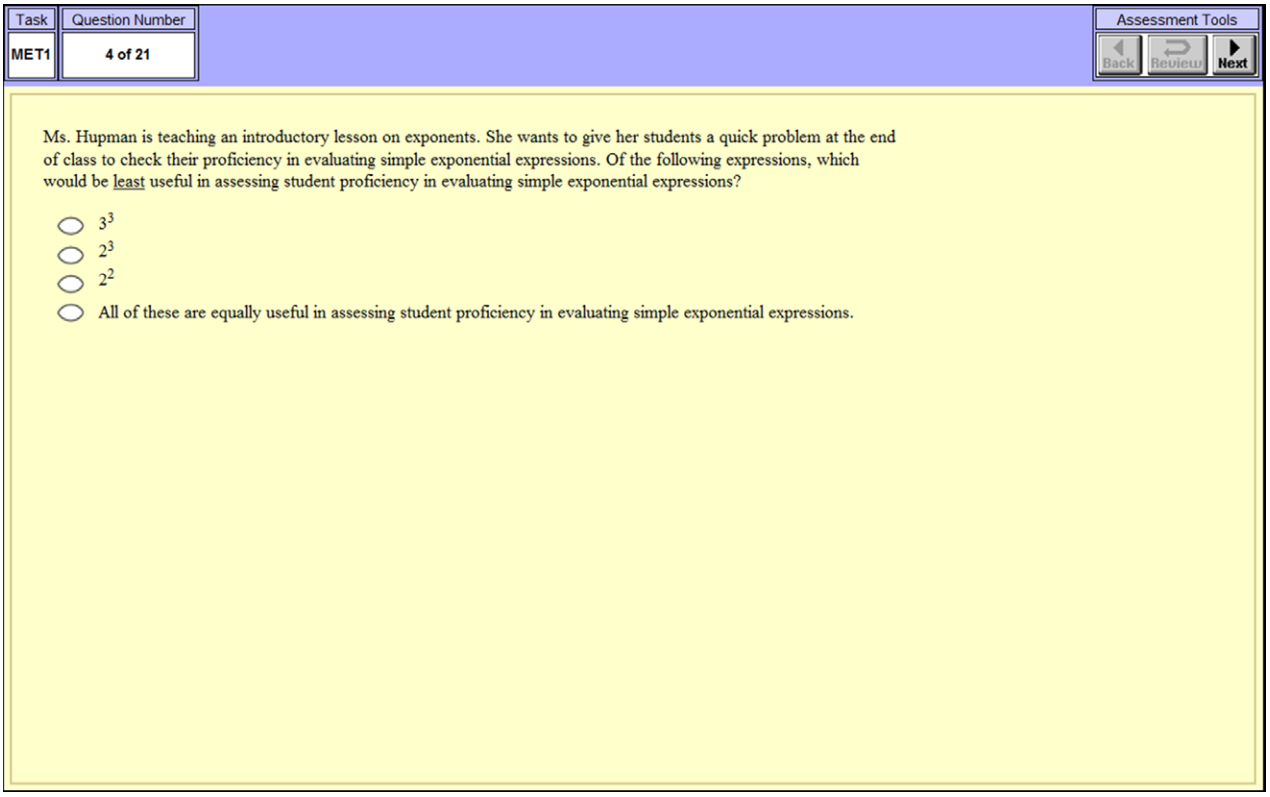



**Figure A4** Question #4.

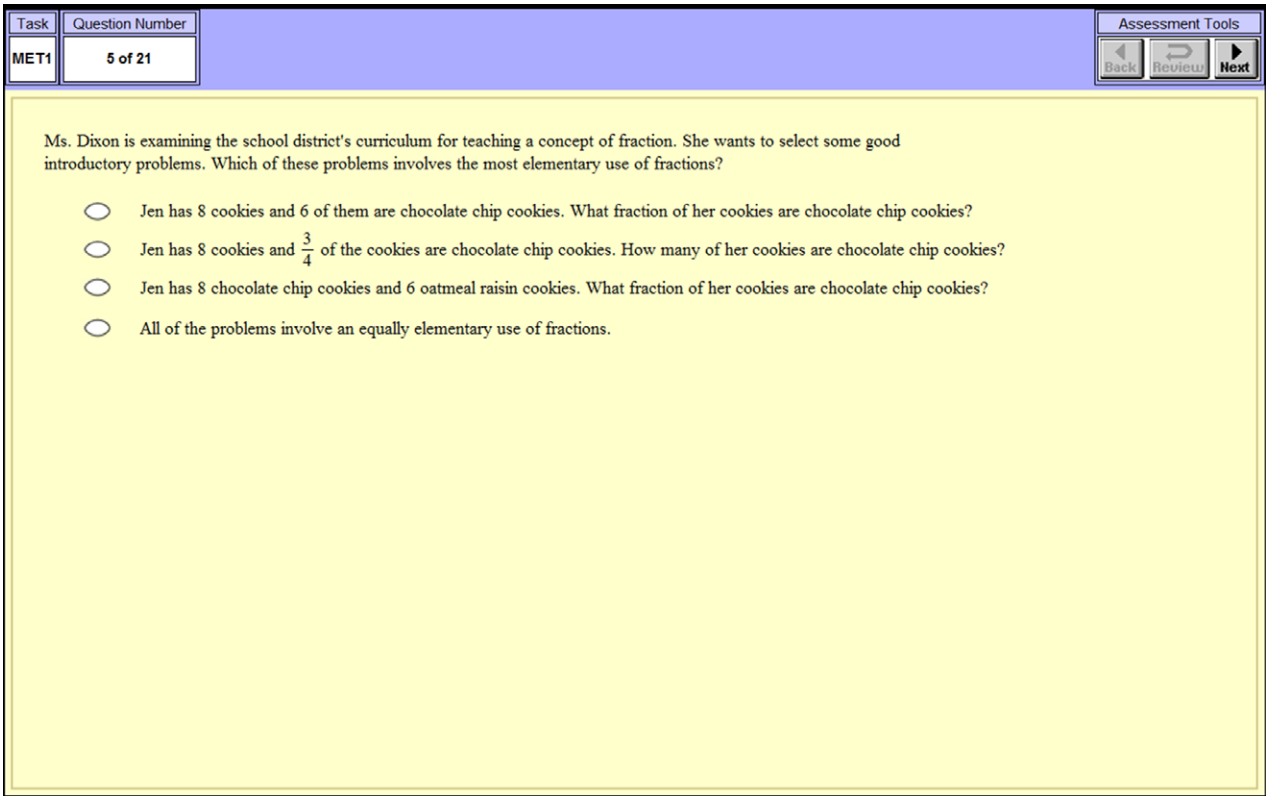



**Figure A5** Question #5.


**Figure A6** Question #6.

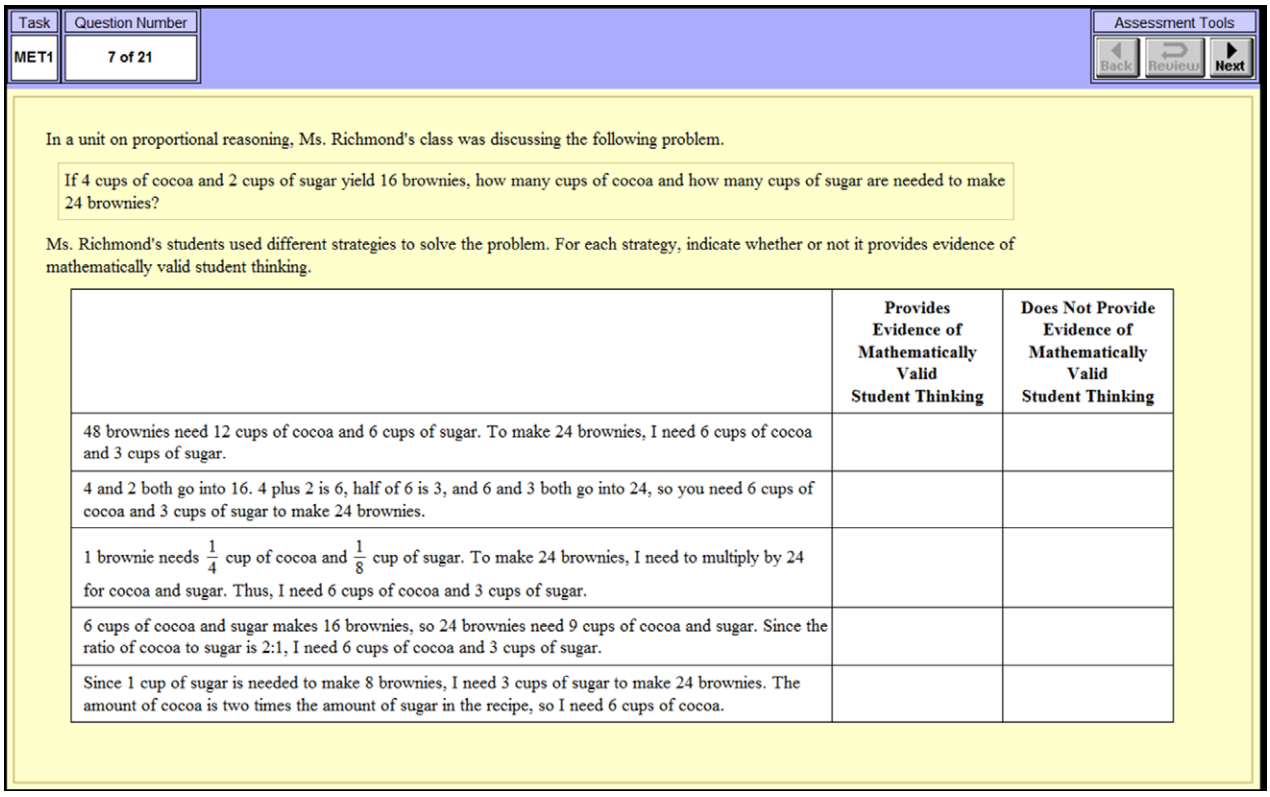



**Figure A7** Question #7.

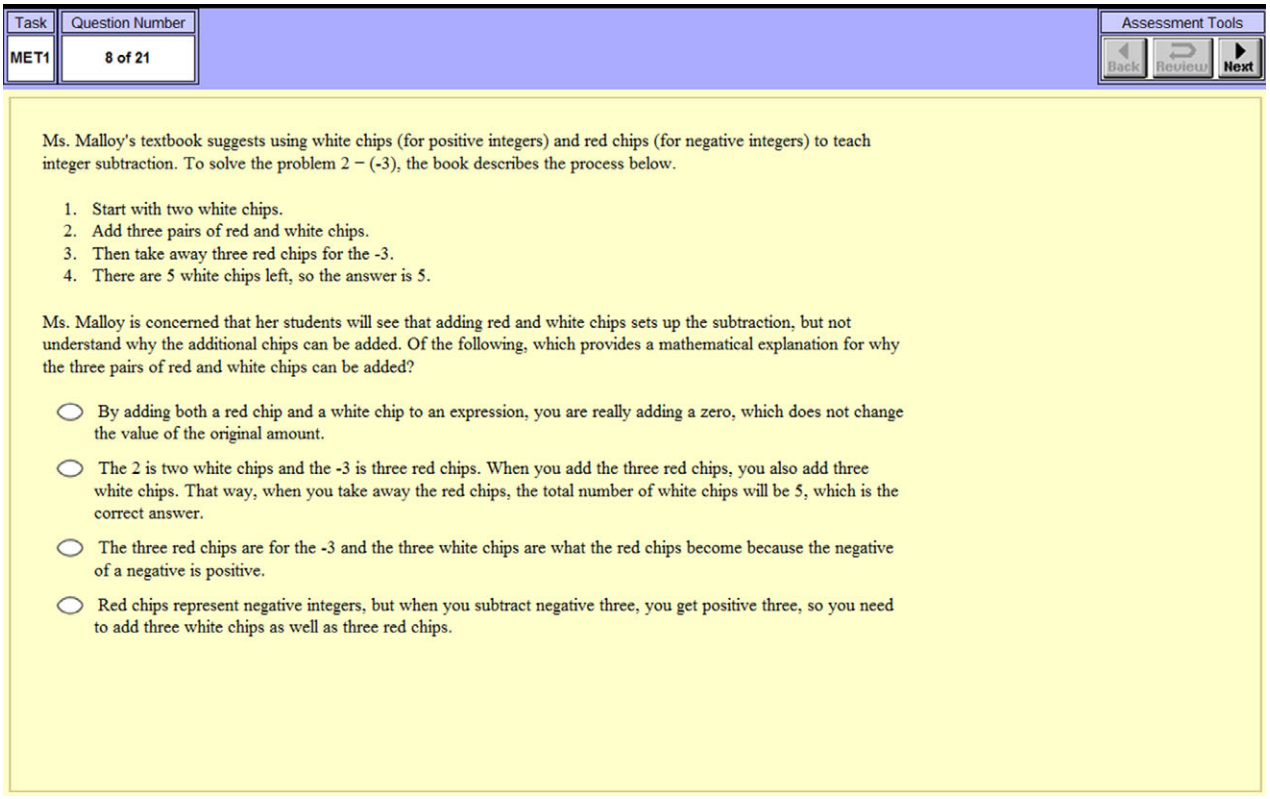
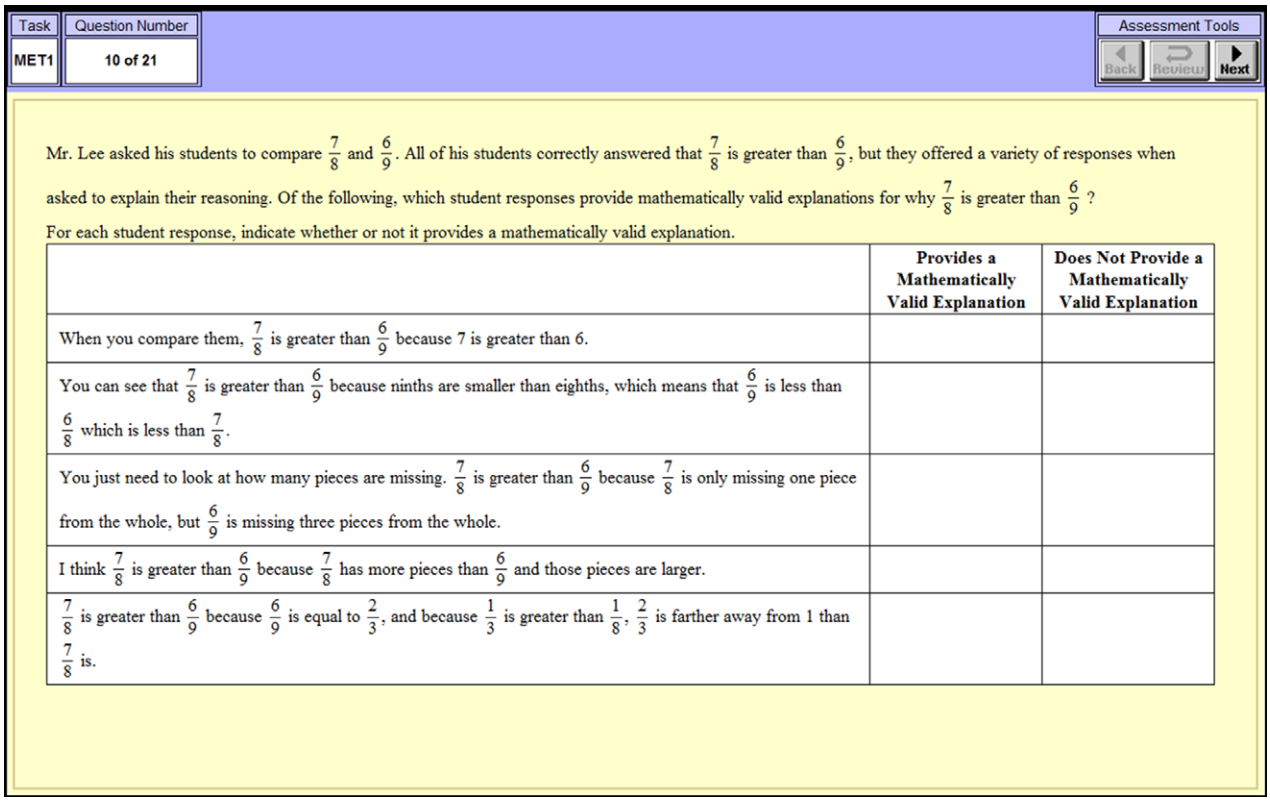



**Figure A8** Question #8.


**Figure A9** Question #9.

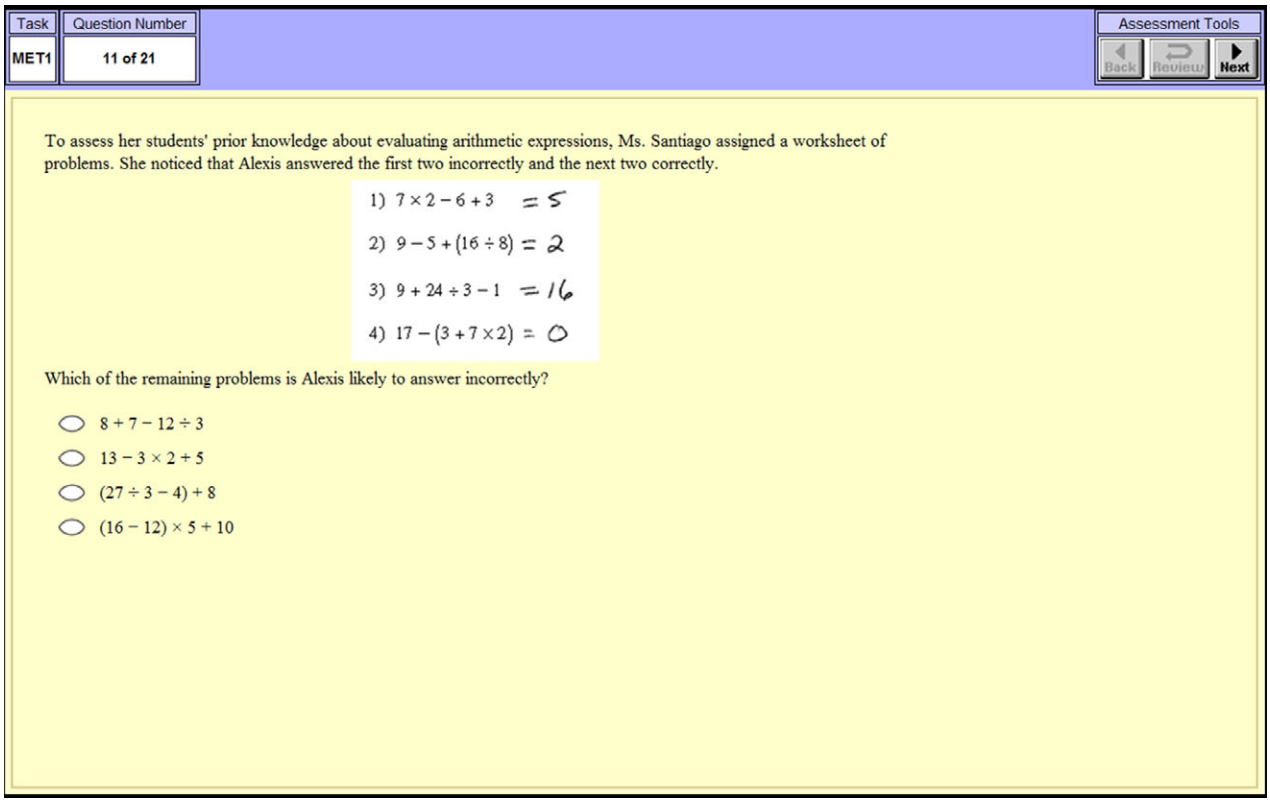



Mr. Lee asked his students to compare $\frac{7}{8}$ and $\frac{6}{9}$. All of his students correctly answered that $\frac{7}{8}$ is greater than $\frac{6}{9}$, but they offered a variety of responses when asked to explain their reasoning. Of the following, which student responses provide mathematically valid explanations for why $\frac{7}{8}$ is greater than $\frac{6}{9}$ ?

For each student response, indicate whether or not it provides a mathematically valid explanation.

| | **Provides a Mathematically Valid Explanation** | **Does Not Provide a Mathematically Valid Explanation** |
|---|---|---|
| When you compare them, $\frac{7}{8}$ is greater than $\frac{6}{9}$ because 7 is greater than 6. | | |
| You can see that $\frac{7}{8}$ is greater than $\frac{6}{9}$ because ninths are smaller than eighths, which means that $\frac{6}{9}$ is less than $\frac{6}{8}$ which is less than $\frac{7}{8}$. | | |
| You just need to look at how many pieces are missing. $\frac{7}{8}$ is greater than $\frac{6}{9}$ because $\frac{7}{8}$ is only missing one piece from the whole, but $\frac{6}{9}$ is missing three pieces from the whole. | | |
| I think $\frac{7}{8}$ is greater than $\frac{6}{9}$ because $\frac{7}{8}$ has more pieces than $\frac{6}{9}$ and those pieces are larger. | | |
| $\frac{7}{8}$ is greater than $\frac{6}{9}$ because $\frac{6}{9}$ is equal to $\frac{2}{3}$, and because $\frac{1}{3}$ is greater than $\frac{1}{8}$, $\frac{2}{3}$ is farther away from 1 than $\frac{7}{8}$ is. | | |

**Figure A10** Question #10.




To assess her students' prior knowledge about evaluating arithmetic expressions, Ms. Santiago assigned a worksheet of problems. She noticed that Alexis answered the first two incorrectly and the next two correctly.

1) $7 \times 2 - 6 + 3 = 5$

2) $9 - 5 + (16 \div 8) = 2$

3) $9 + 24 \div 3 - 1 = 16$

4) $17 - (3 + 7 \times 2) = 0$

Which of the remaining problems is Alexis likely to answer incorrectly?

- ◯ $8 + 7 - 12 \div 3$
- ◯ $13 - 3 \times 2 + 5$
- ◯ $(27 \div 3 - 4) + 8$
- ◯ $(16 - 12) \times 5 + 10$

**Figure A11** Question #11.

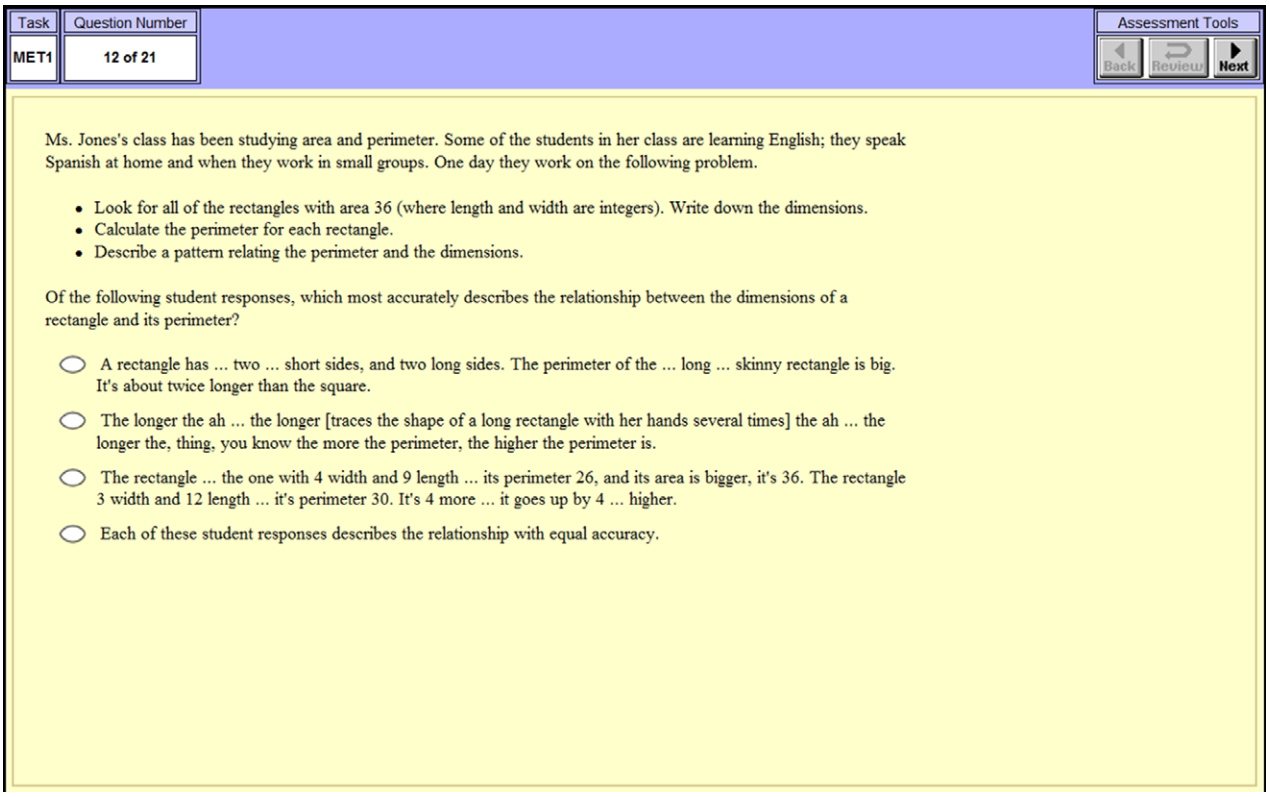



**Figure A12** Question #12.

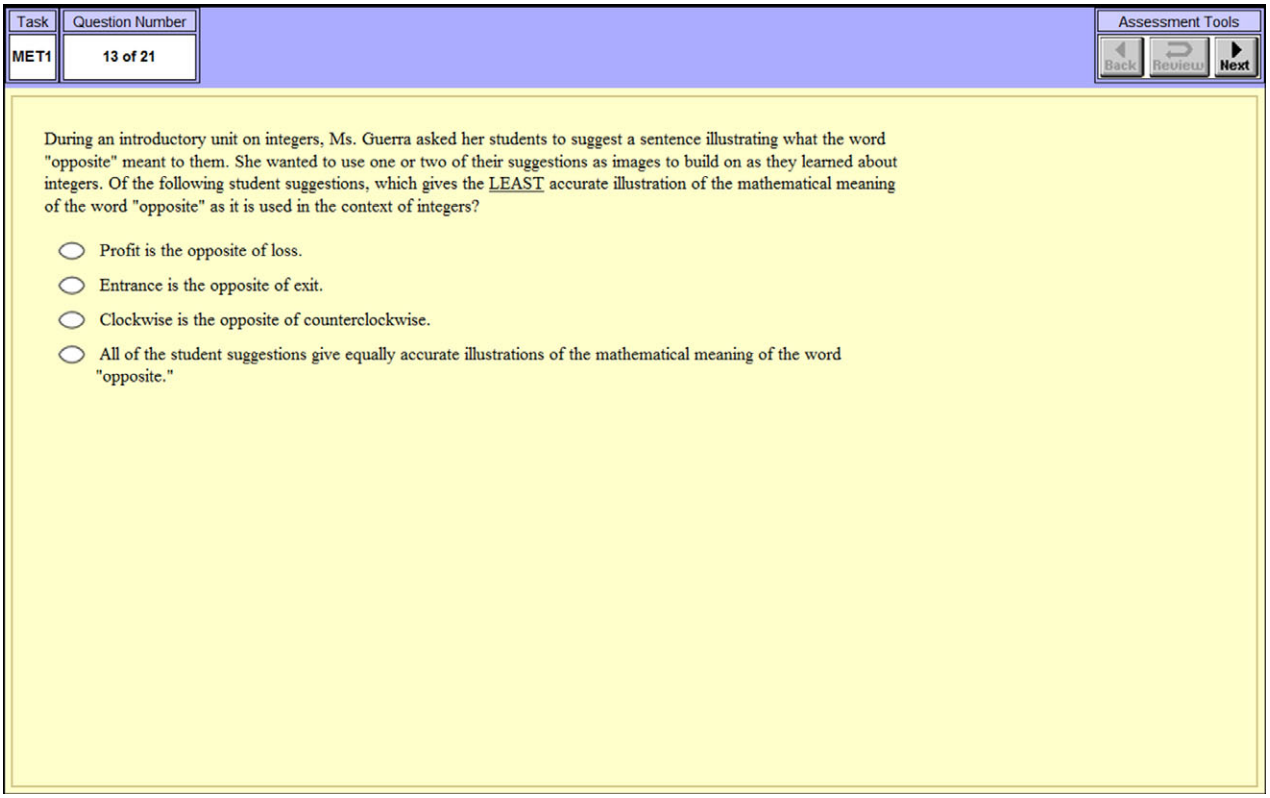



**Figure A13** Question #13.

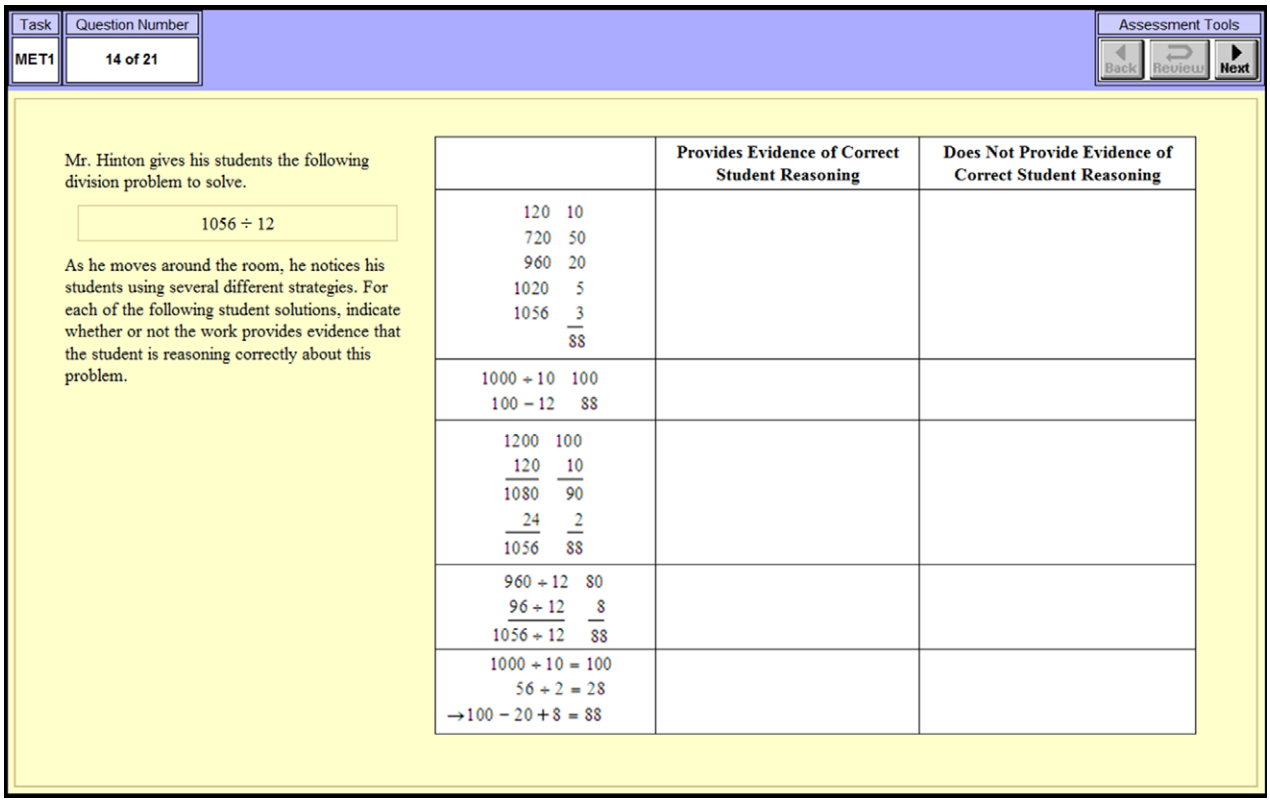



**Figure A14** Question #14.

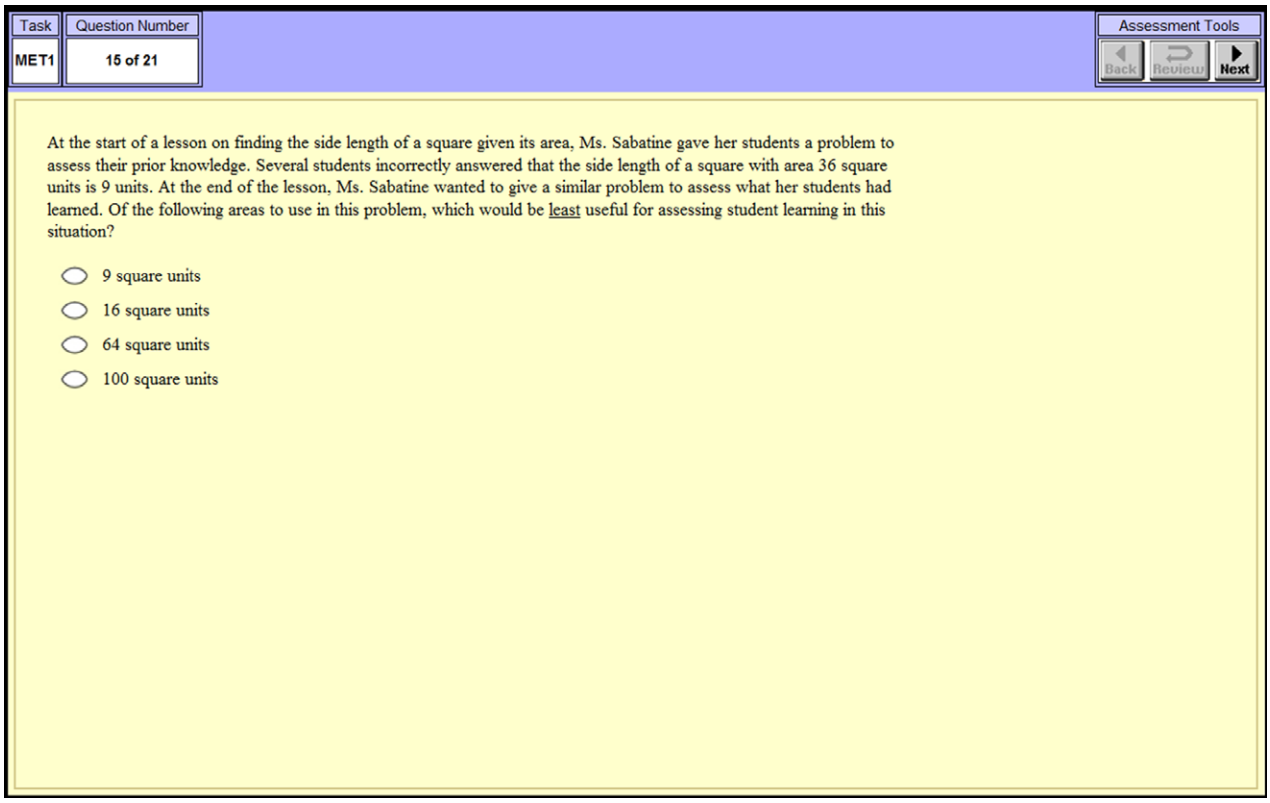



**Figure A15** Question #15.

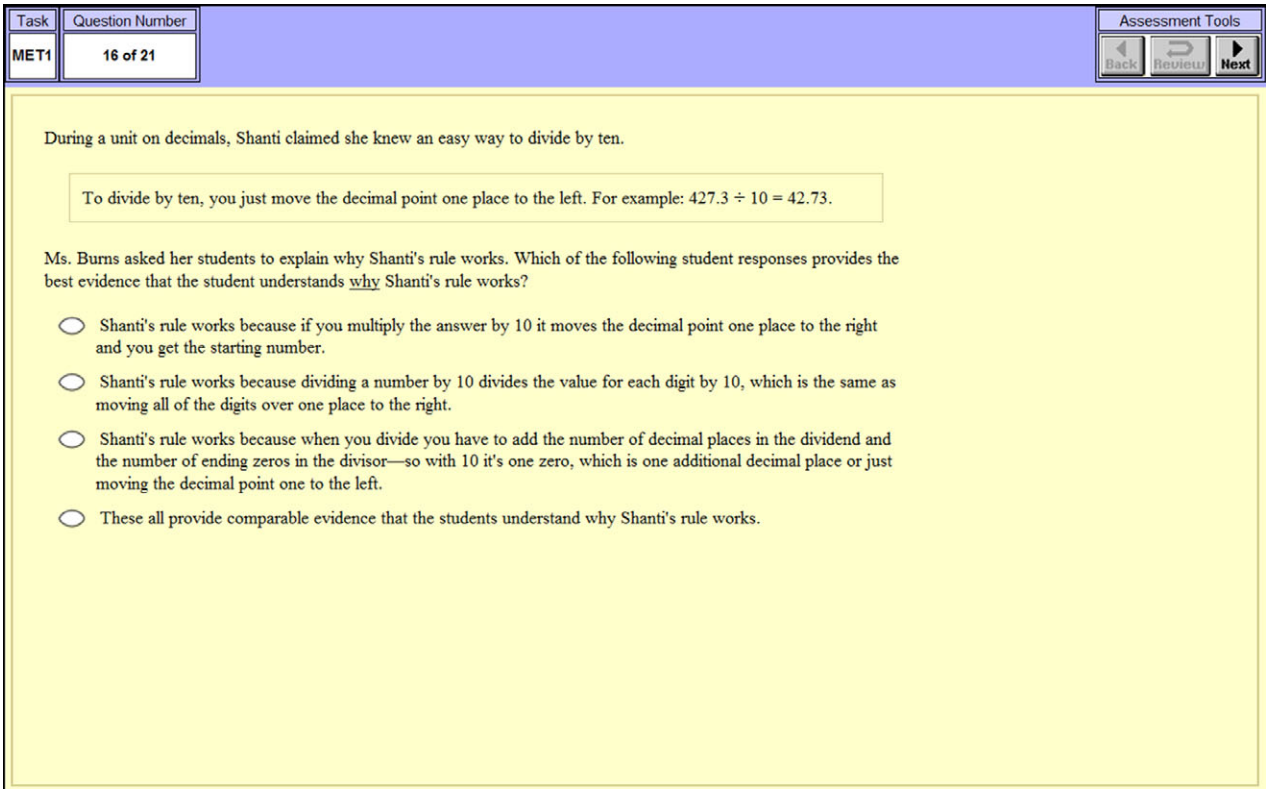



**Figure A16** Question #16.

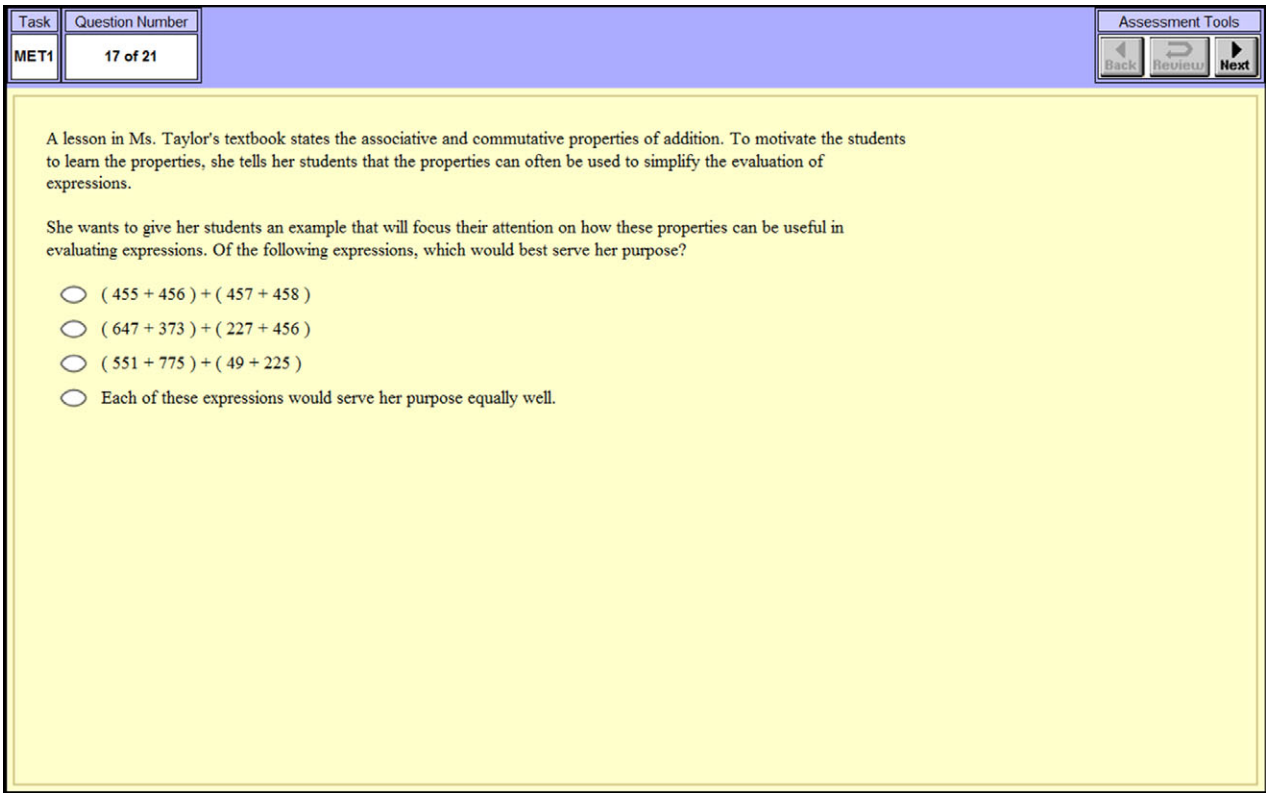



**Figure A17** Question #17.

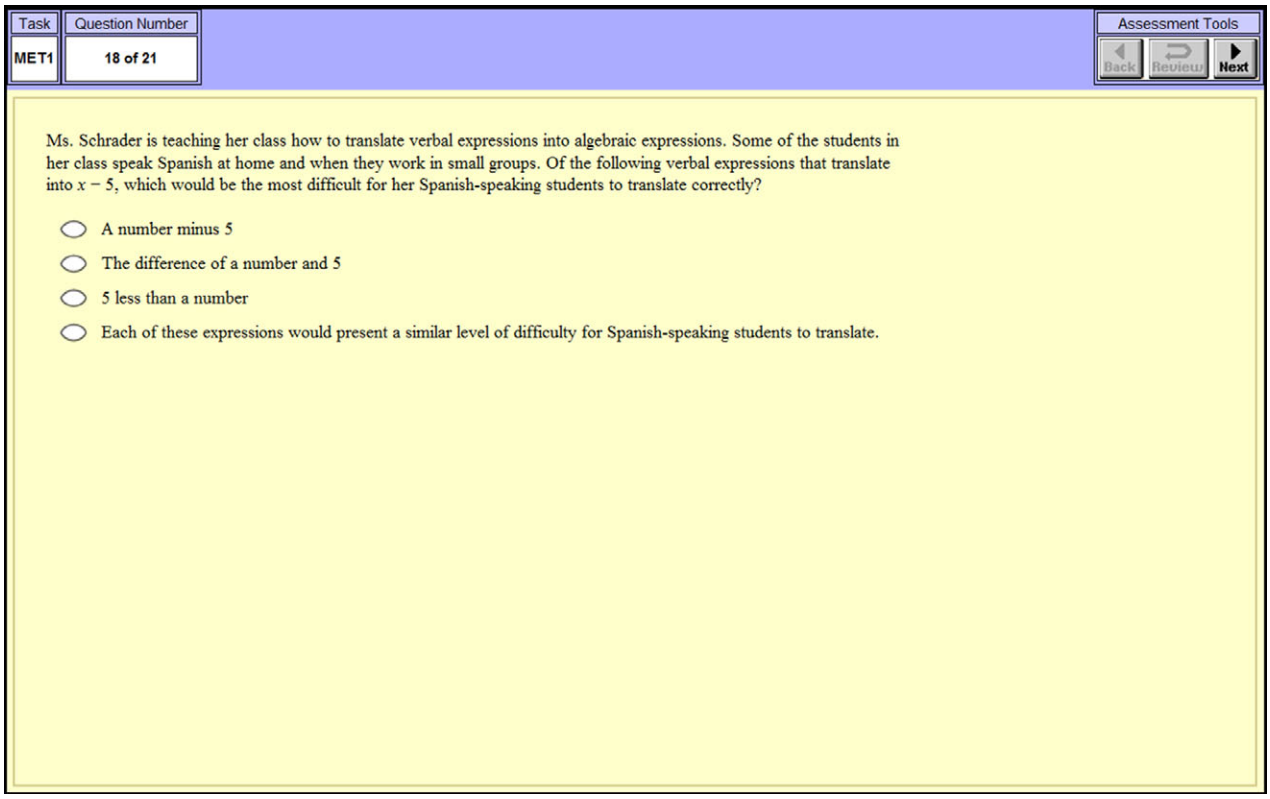



**Figure A18** Question #18.

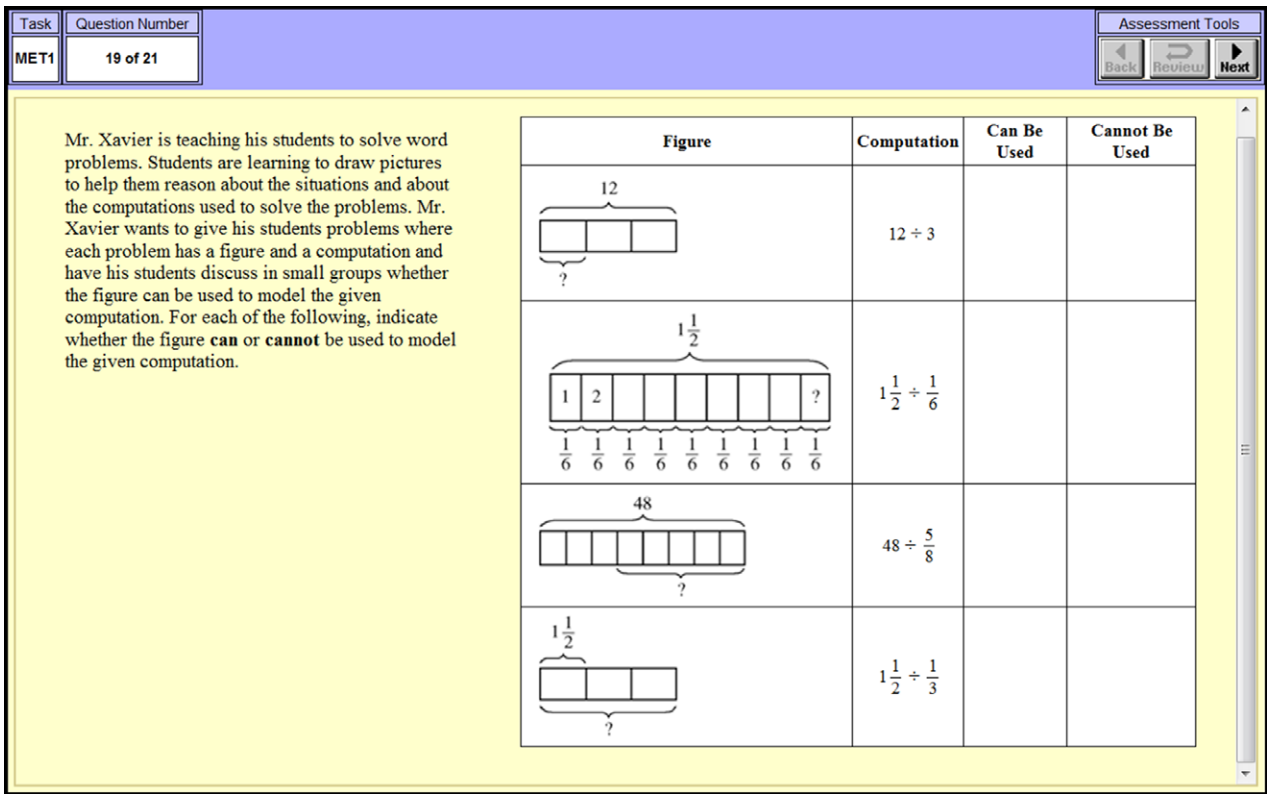



**Figure A19** Question #19.

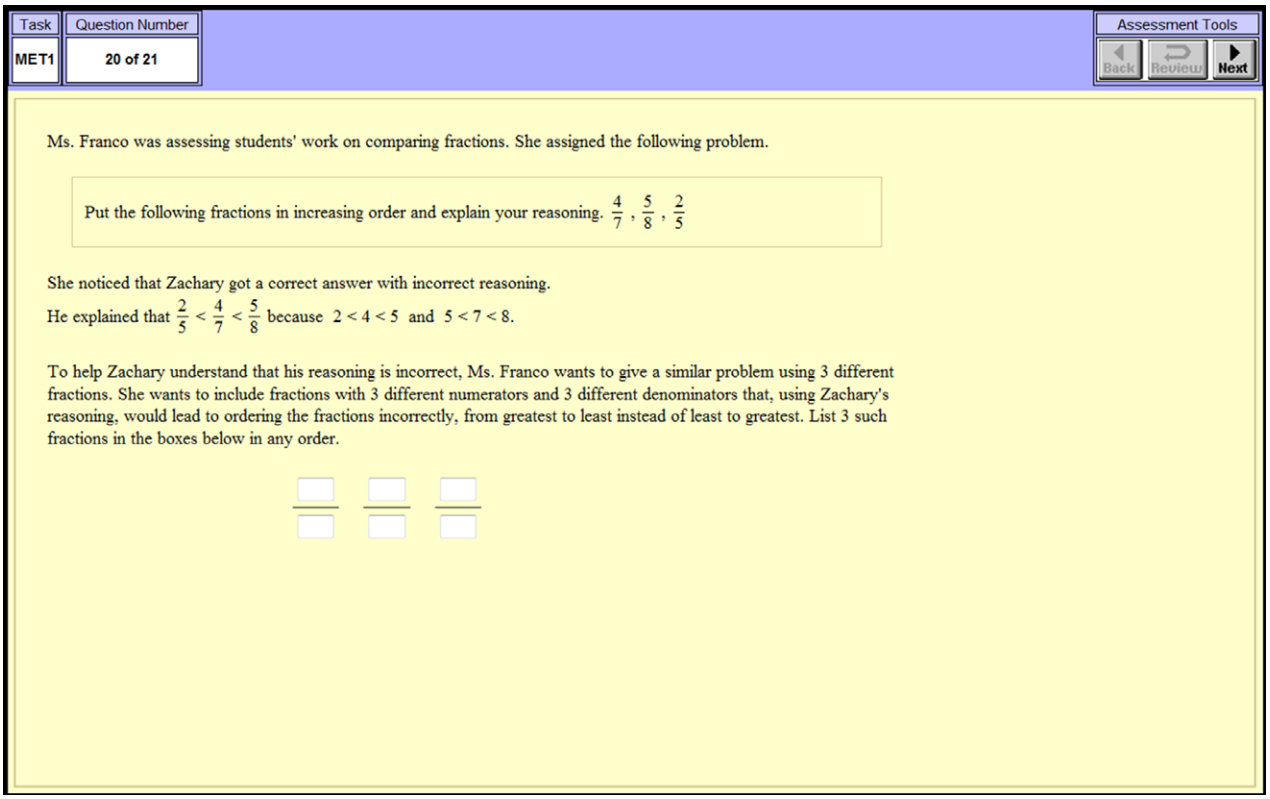



**Figure A20** Question #20.

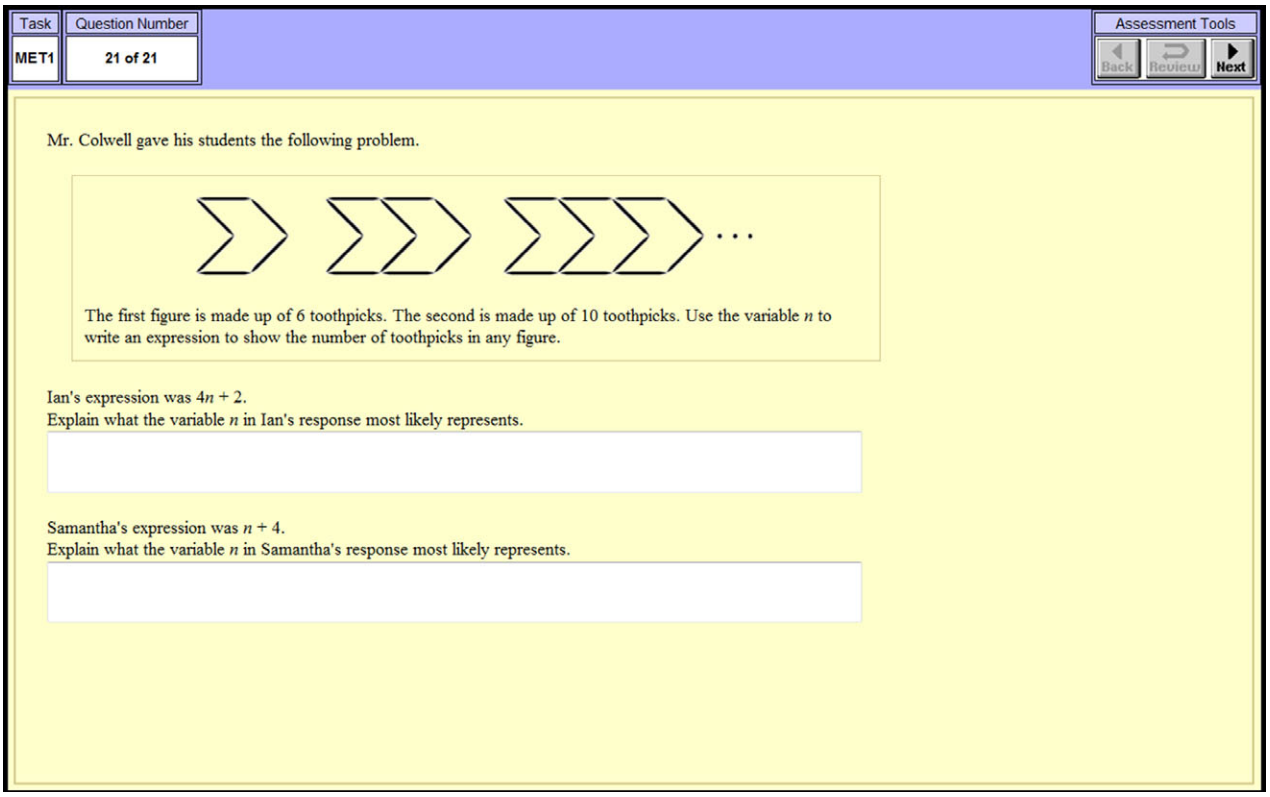



**Figure A21** Question #21.

## Appendix B

## English Language Arts (ELA) 4–6 Assessment

**Table B1** Item/Question Sequence Number for English Language Arts (ELA) 4–6 Assessment

| Item/question sequence number | Answer key | Item/question sequence number | Answer key |
|---|---|---|---|
| 1 | B | 12d | Would not |
| 2 | A | 12e | Would |
| 3 | C | 13 | C |
| 4 | B | 14 | D |
| 5 | A | 15 | B |
| 6a | Yes | 16a | Effective |
| 6b | Likely | 16b | Effective |
| 6c | Likely | 16c | Effective |
| 6d | Not likely | 16d | Not effective |
| 6e | Likely | 16e | Not effective |
| 6f | Not likely | 17 | B |
| 7 | C | 18 | Open-ended response |
| 8a | Will | 19a | Will help |
| 8b | Will not | 19b | Will help |
| 8c | Will | 19c | Will not help |
| 8d | Will not | 19d | Will not help |
| 9 | Accurate | 20 | B |
| 10 | D | 21 | C |
| 11a | Accurate | 22 | Open-ended response |
| 11b | Accurate | 23a | Does not describe |
| 11c | Not accurate | 23b | Does not describe |
| 11d | Accurate | 23c | Describes |
| 11e | Not accurate | 23d | Describes |
| 11f | Not accurate | 24 | A |
| 12a | Would not | 25 | D |
| 12b | Would | 26 | B |
| 12c | Would | | |

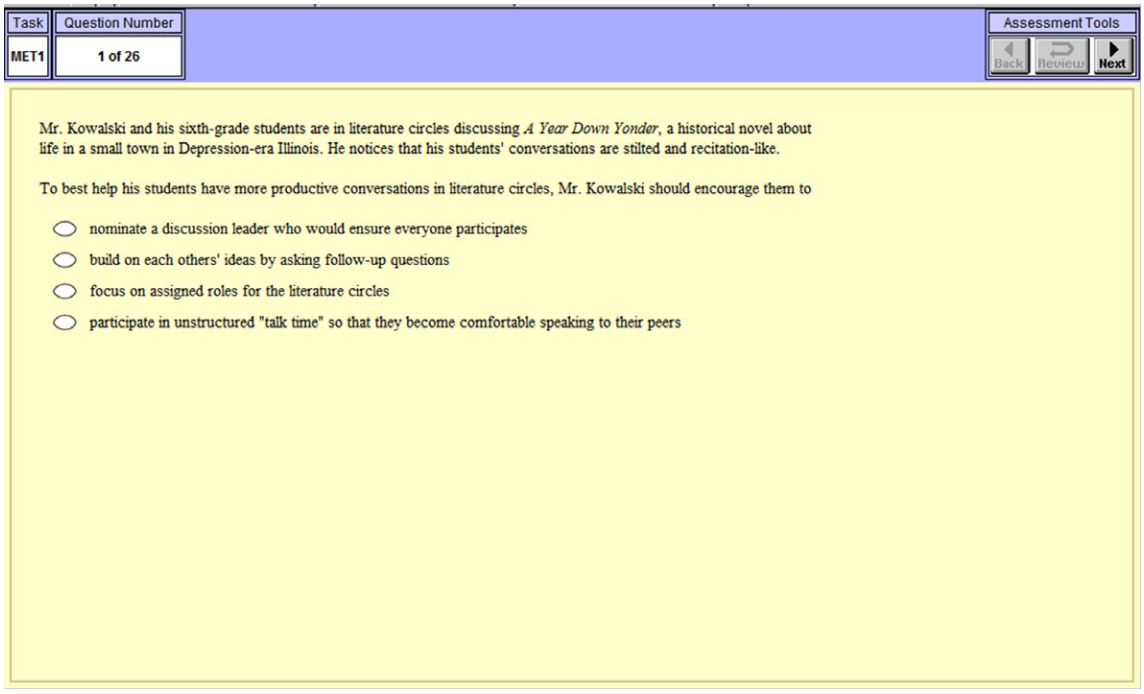



**Figure B1** Question #1.

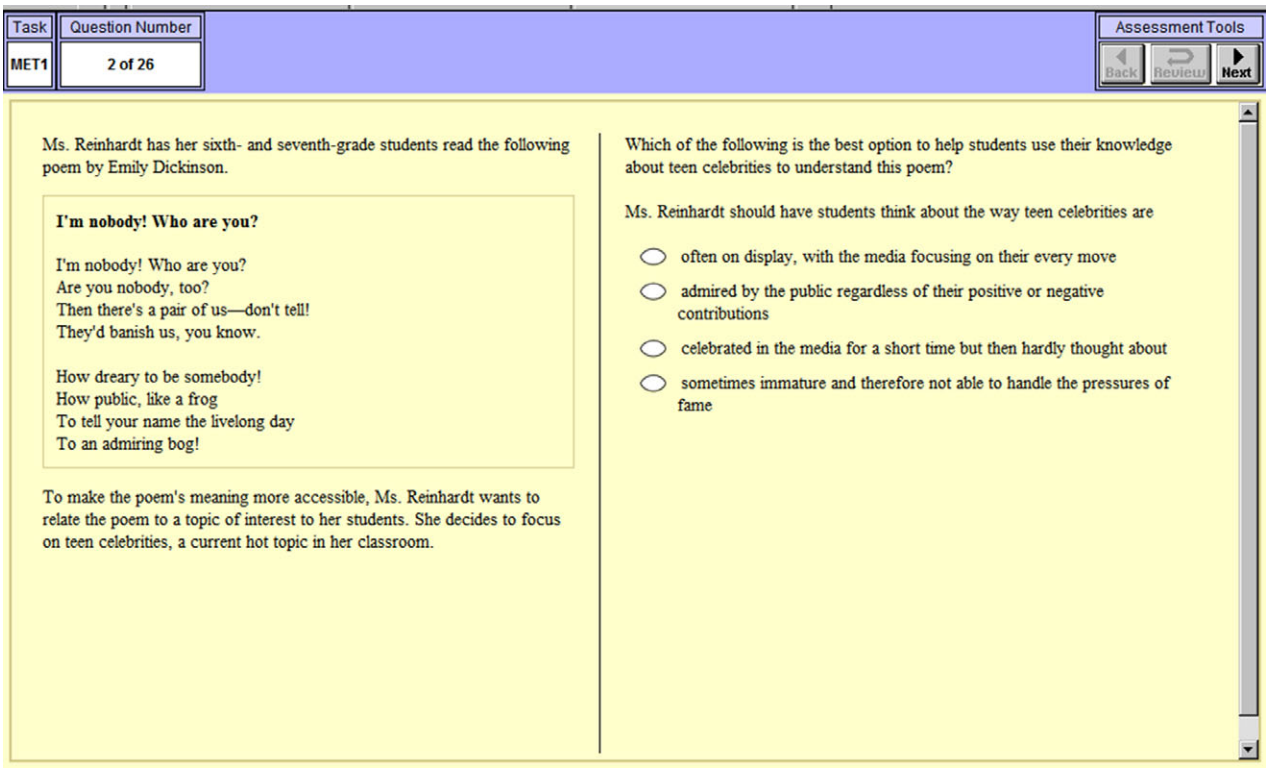



**Figure B2** Question #2. *I'm Nobody by* Emily Dickinson. Reprinted by permission of Dover Publications.

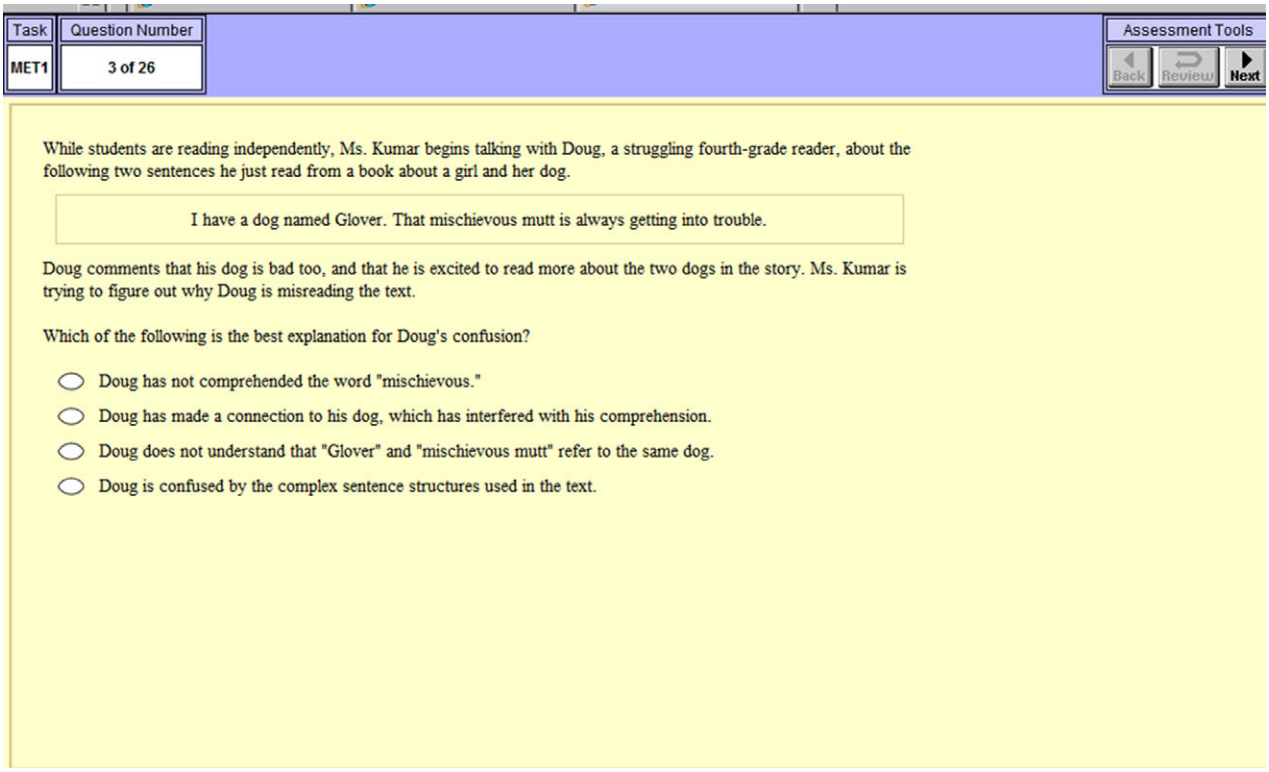



**Figure B3** Question #3.

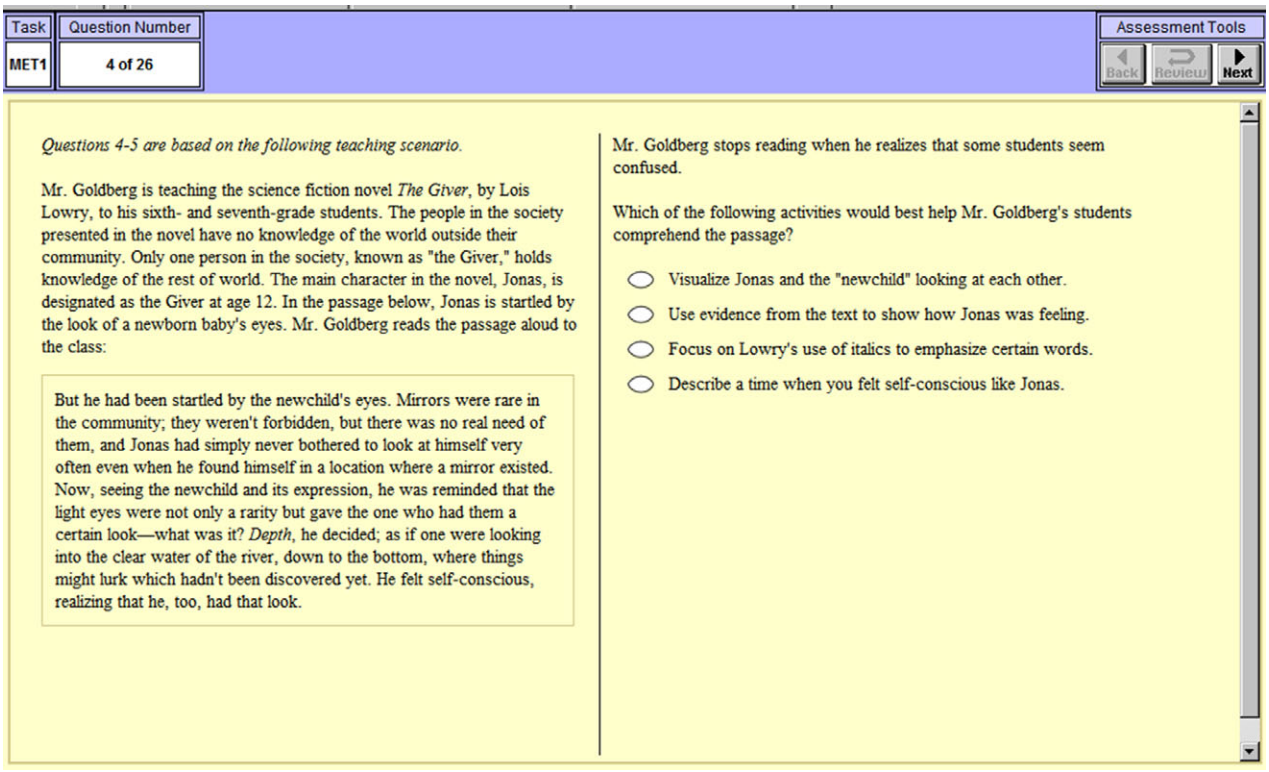



**Figure B4** Question #4.

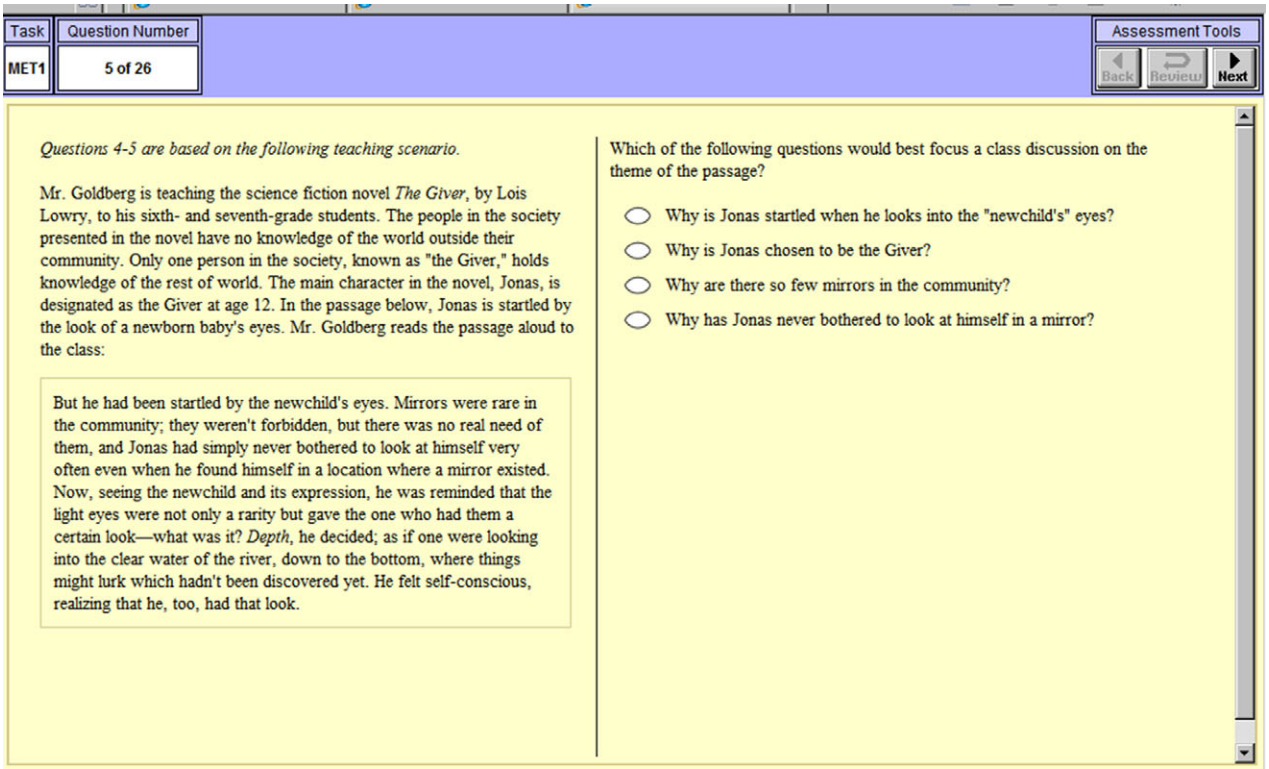



**Figure B5** Question #5.

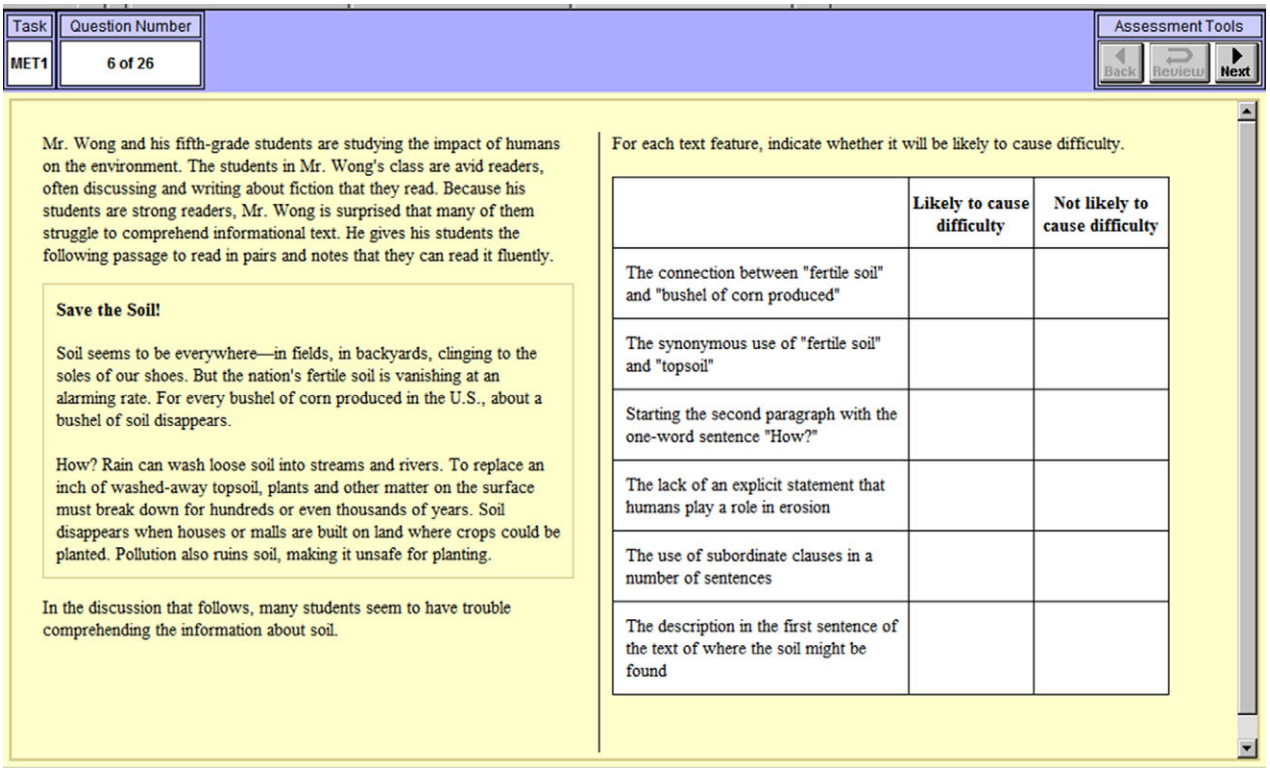



**Figure B6** Question #6. Adapted from *TIME for Kids Magazine,* 2/9/1999 © 1999 Time Inc. Used under license. TIME for Kids and Time Inc. are not affiliated with, and do not endorse, products or services of Educational Testing Service.

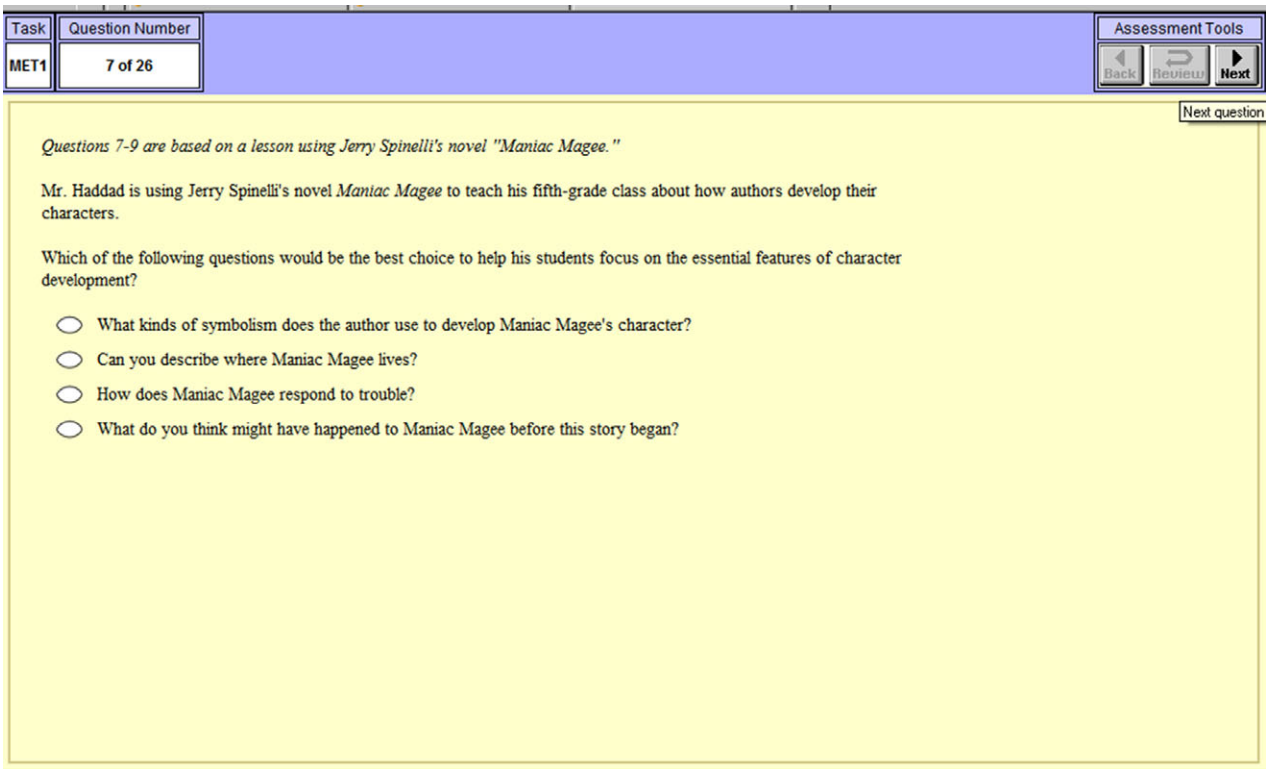



**Figure B7** Question #7.

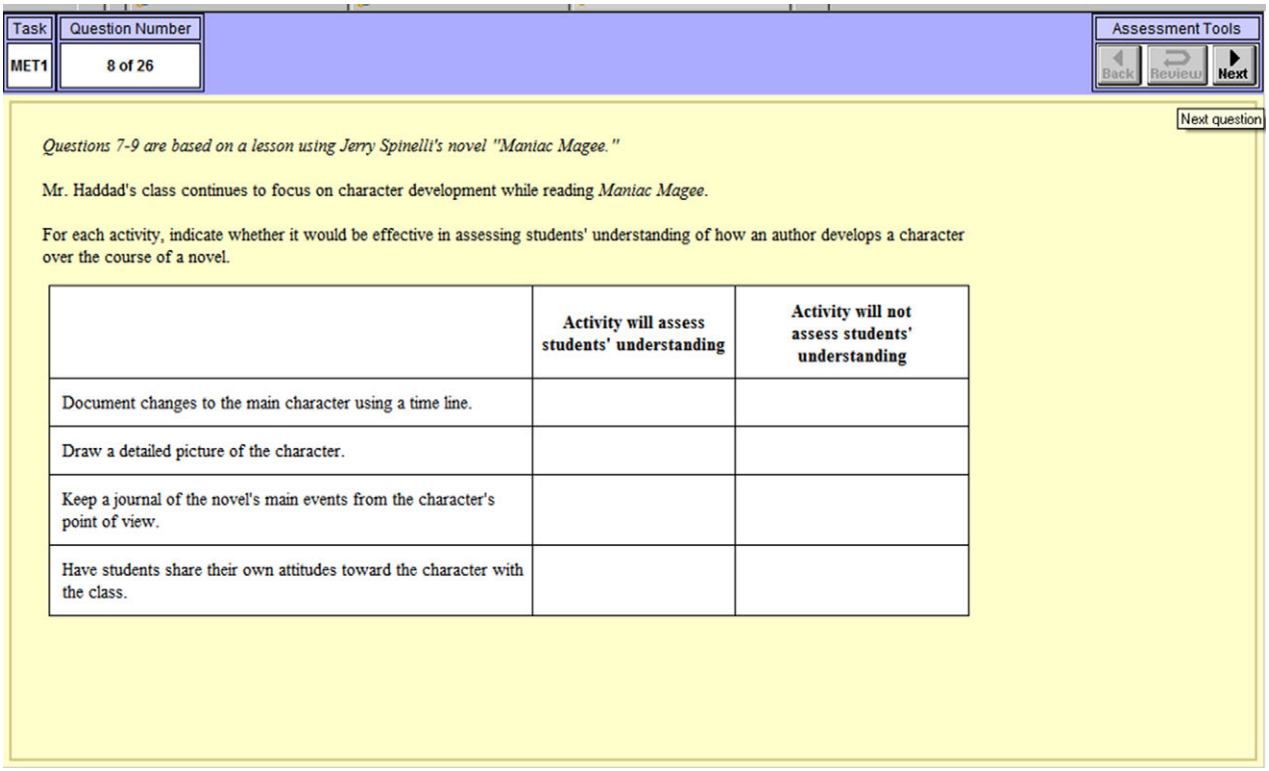



**Figure B8** Question #8.

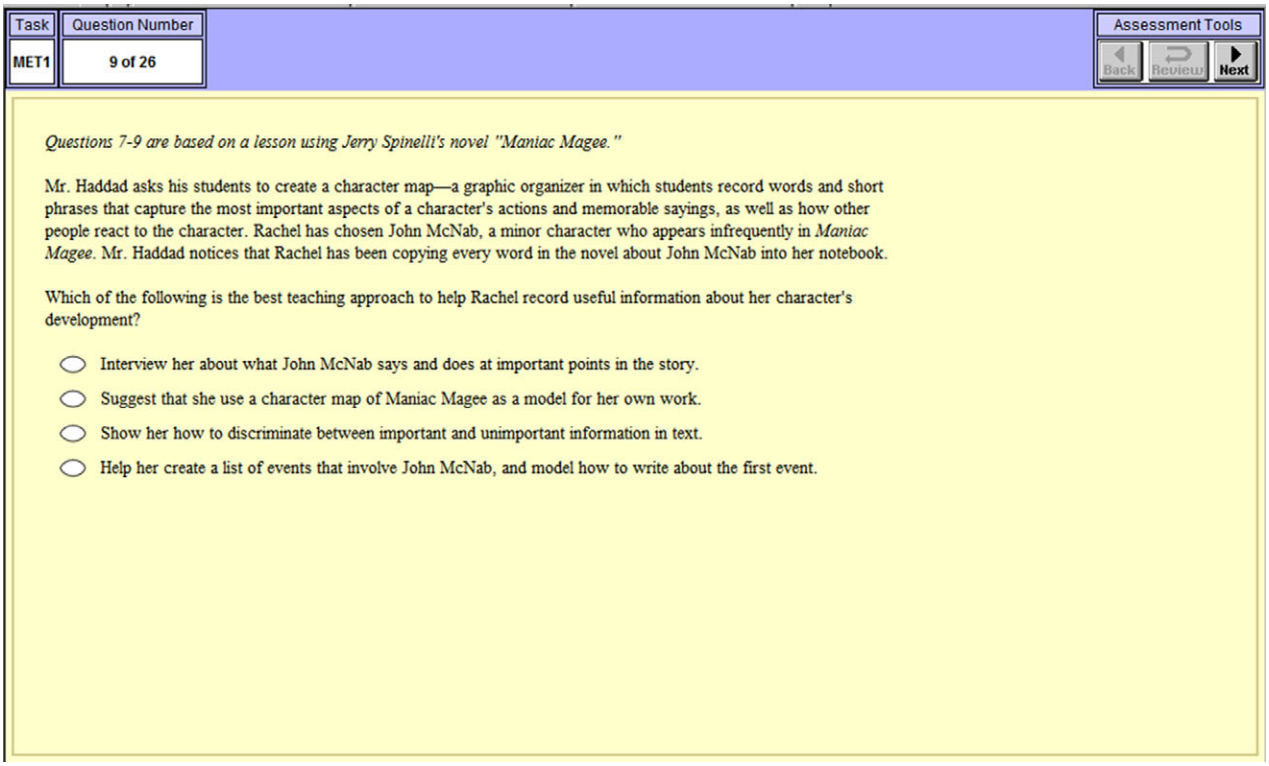



**Figure B9** Question #9.

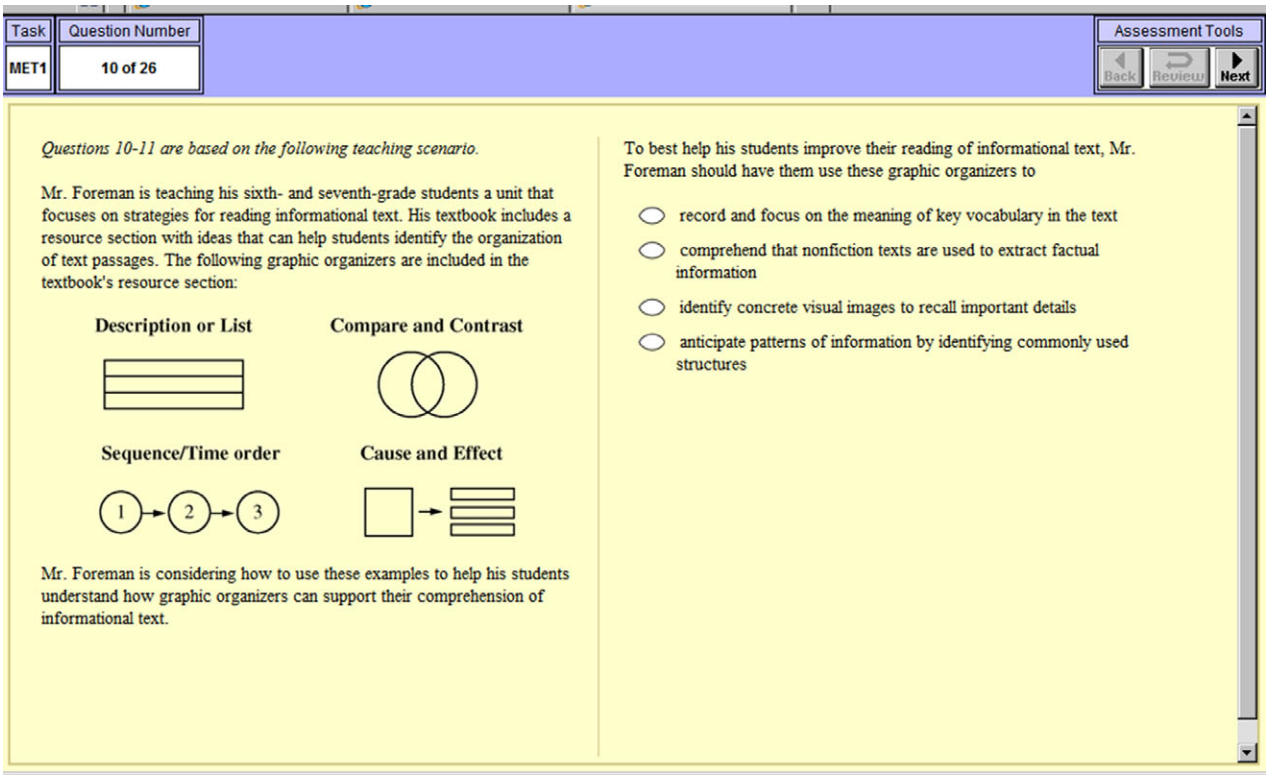



**Figure B10** Question #10.

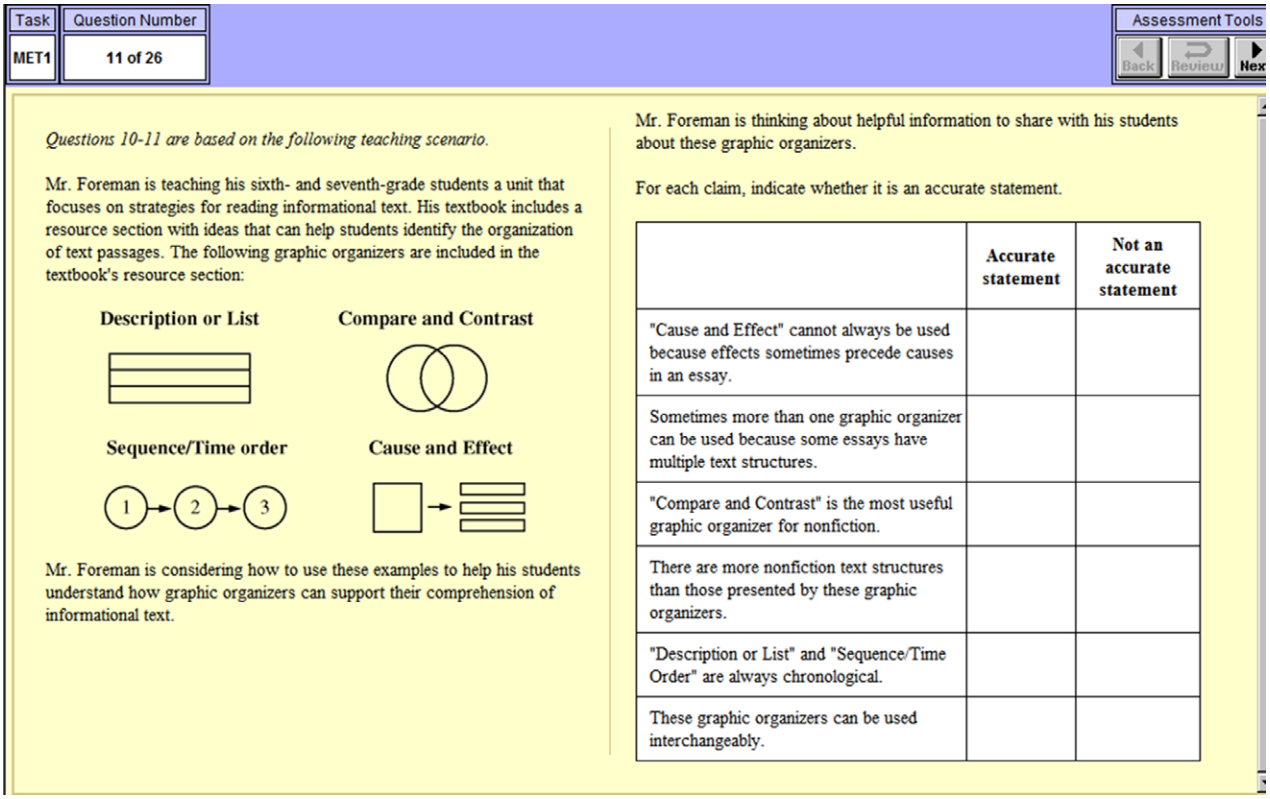



**Figure B11** Question #11.

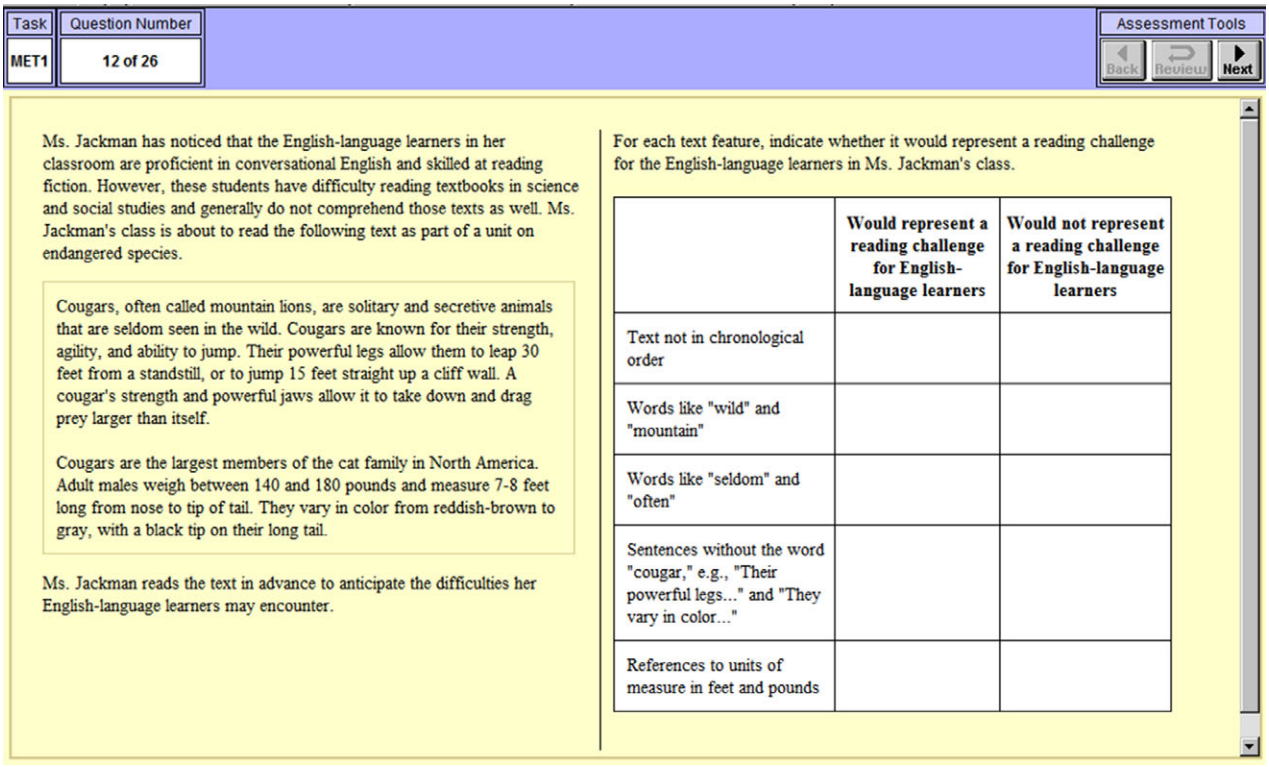



**Figure B12** Question #12.

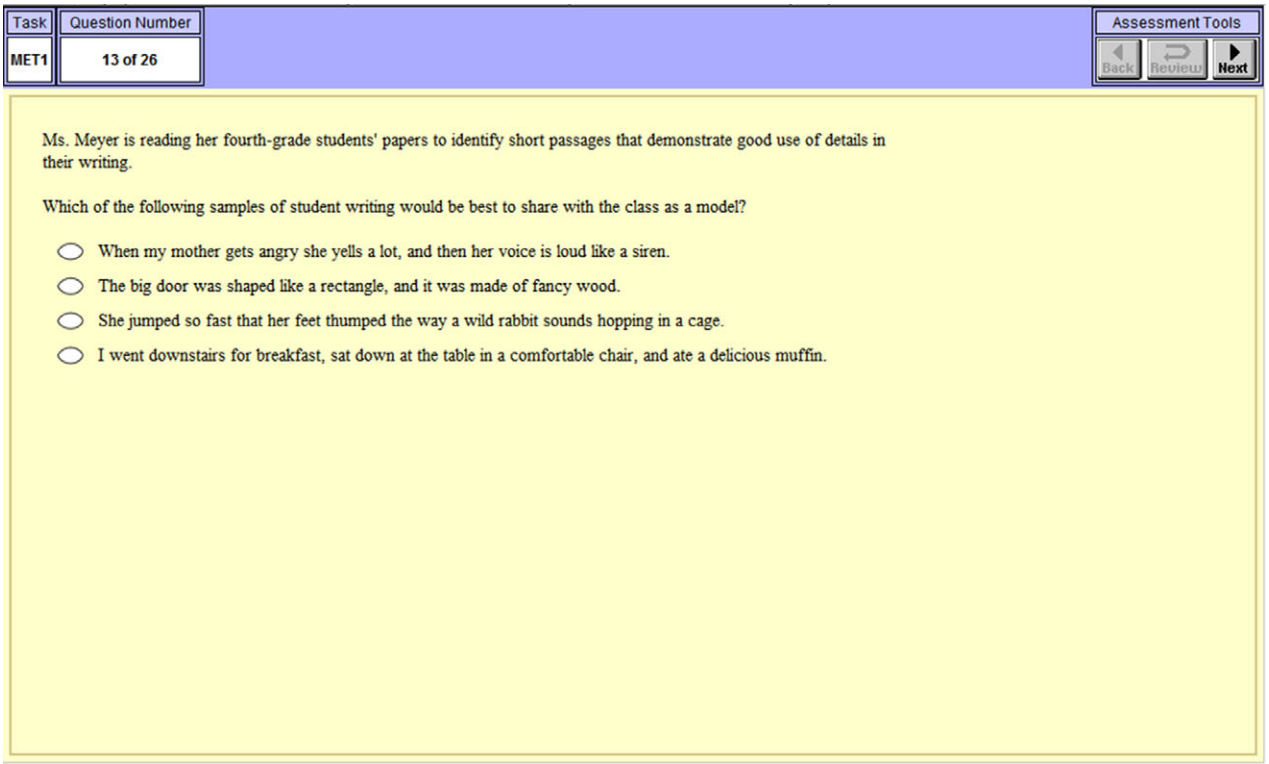



**Figure B13** Question #13.

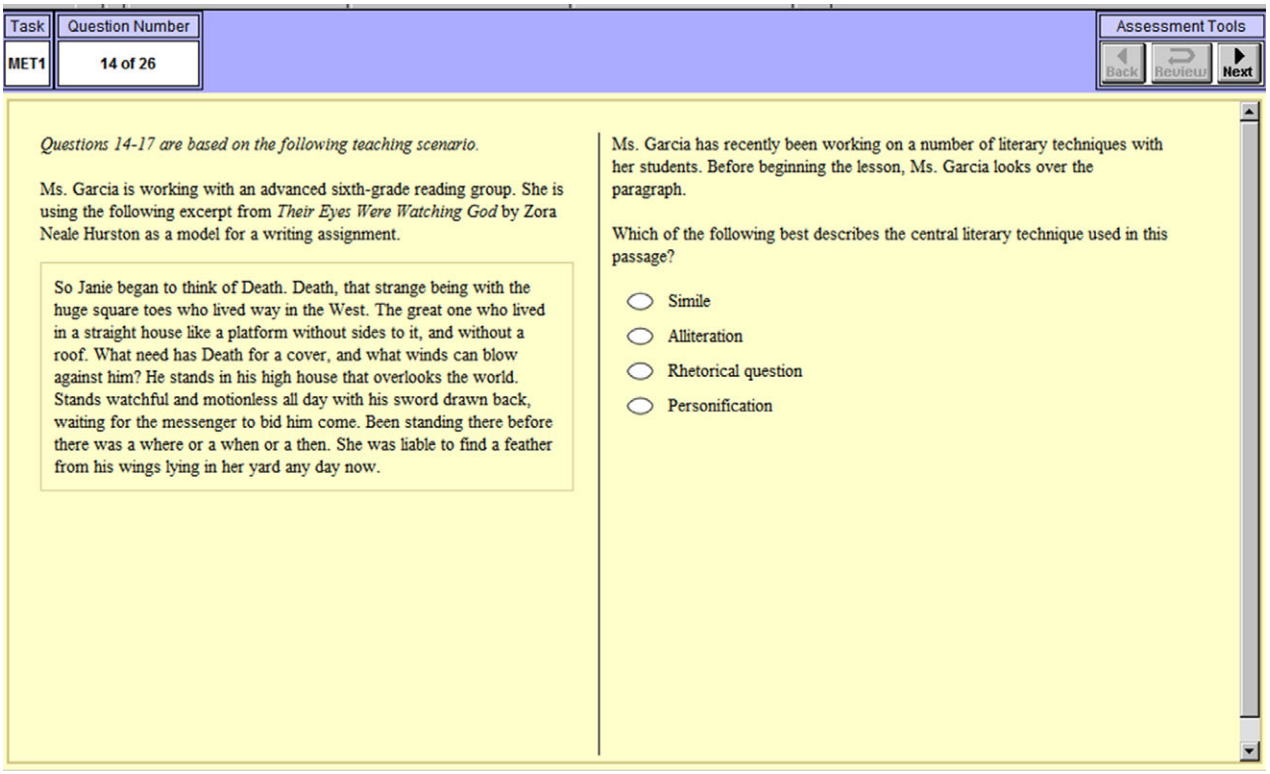



**Figure B14** Question #14.

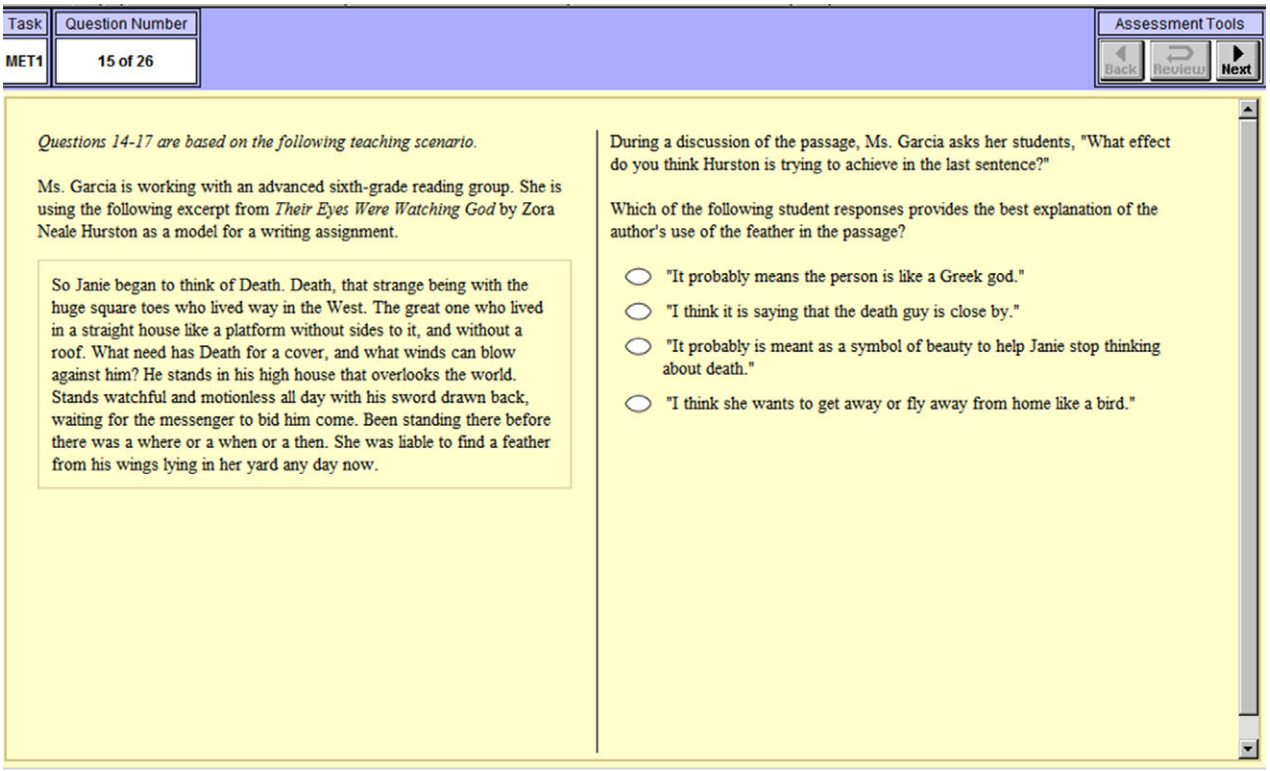



**Figure B15** Question #15.

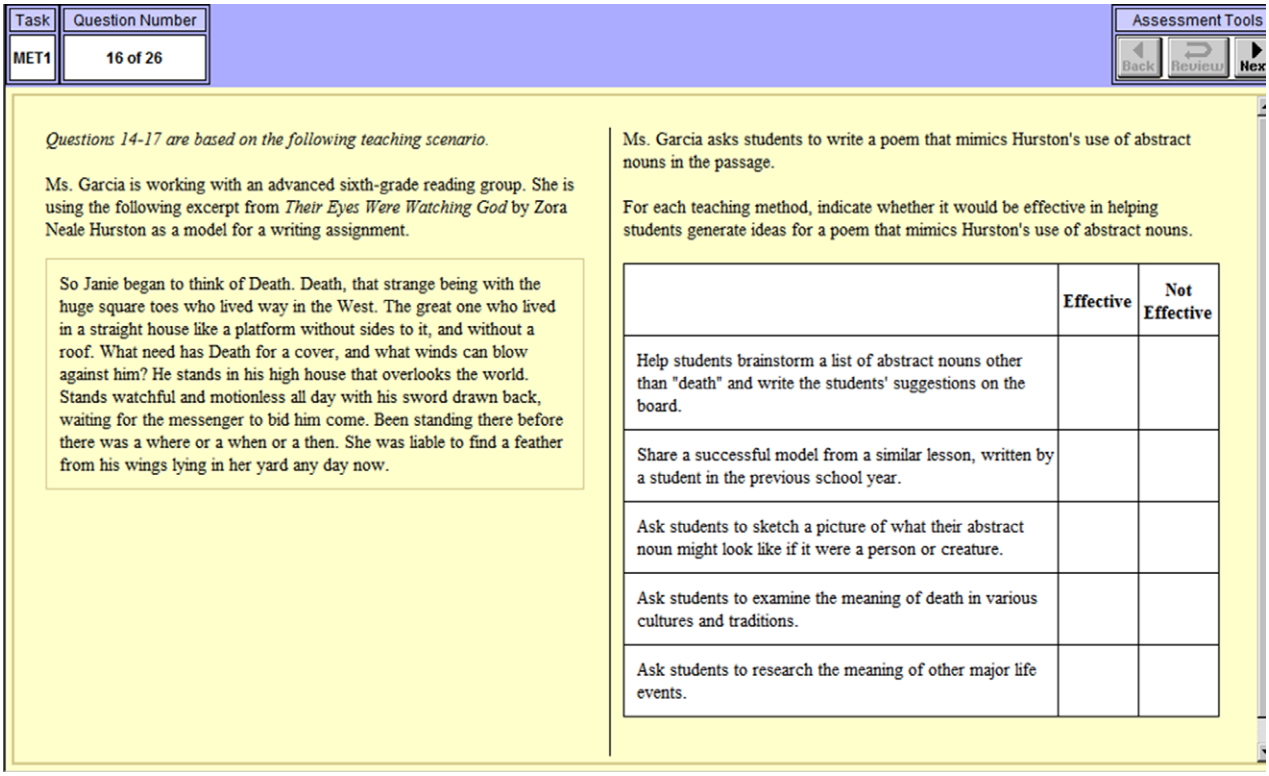



**Figure B16** Question #16.

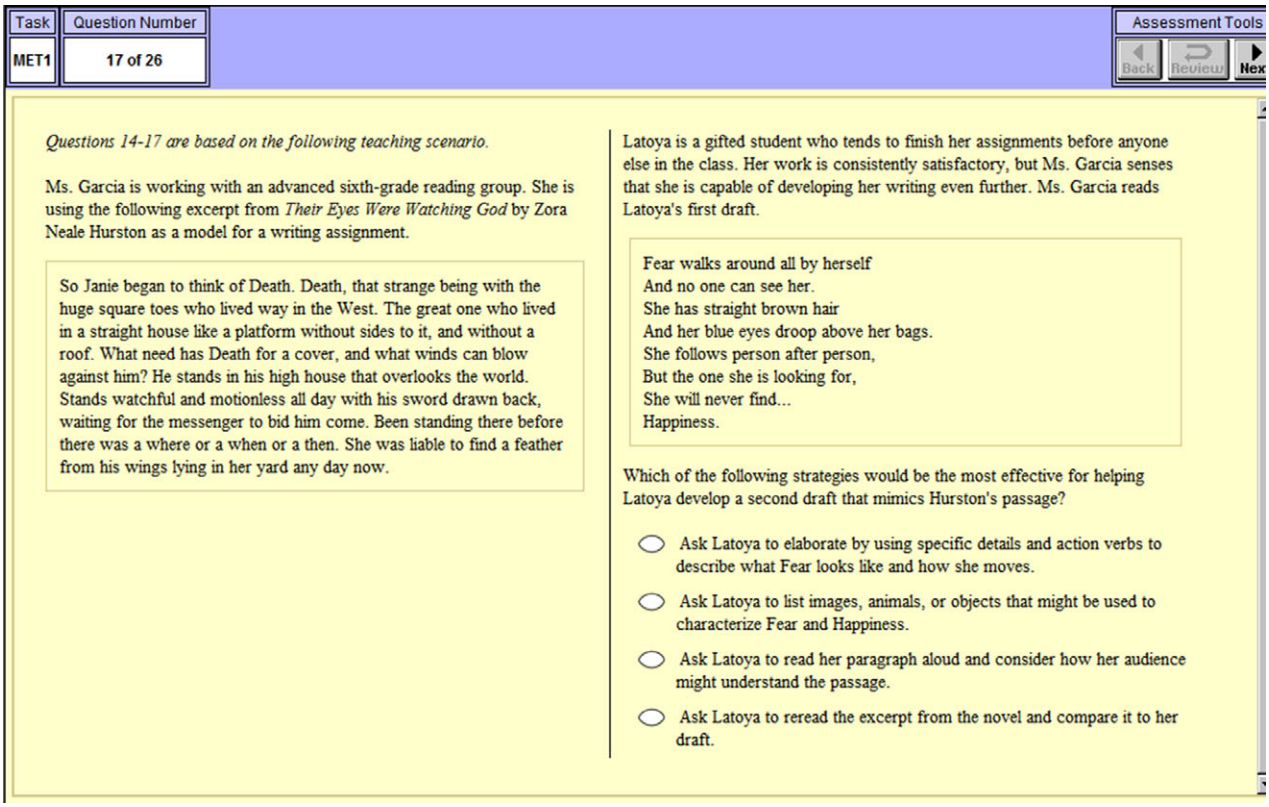



**Figure B17** Question #17.

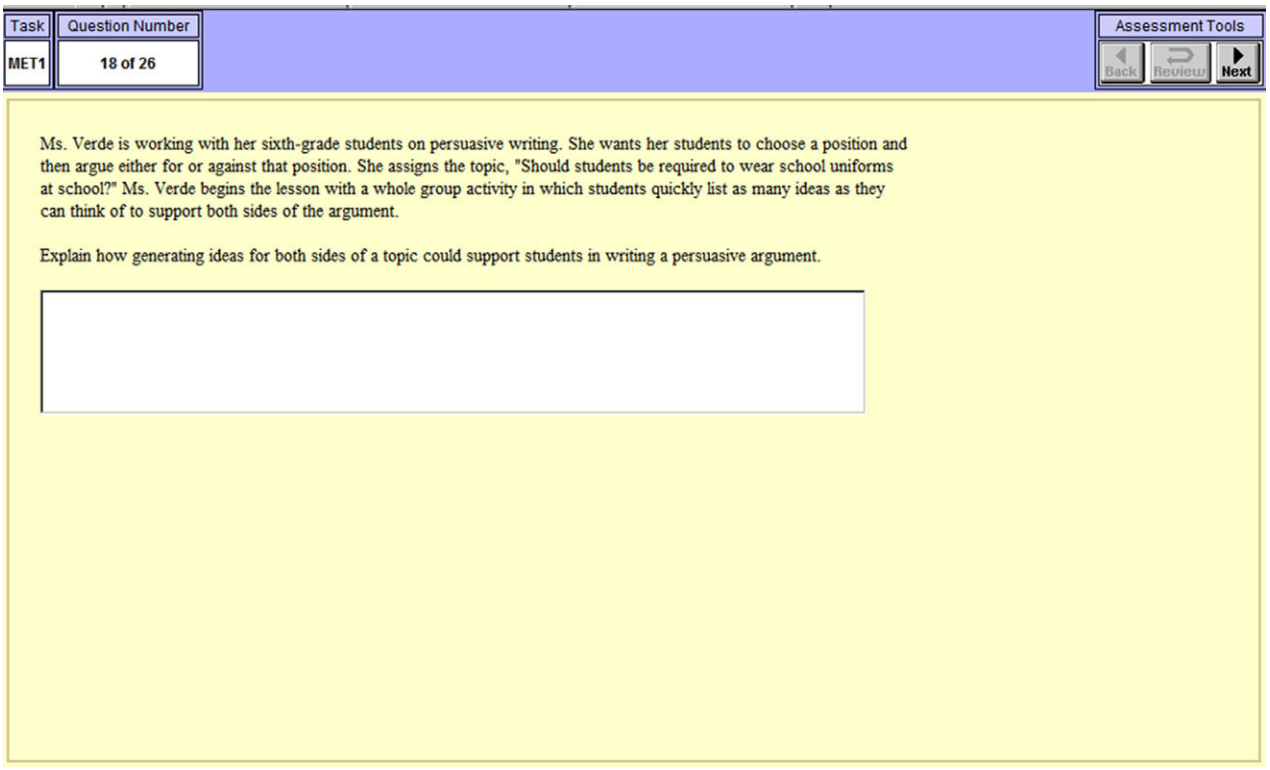



**Figure B18** Question #18.

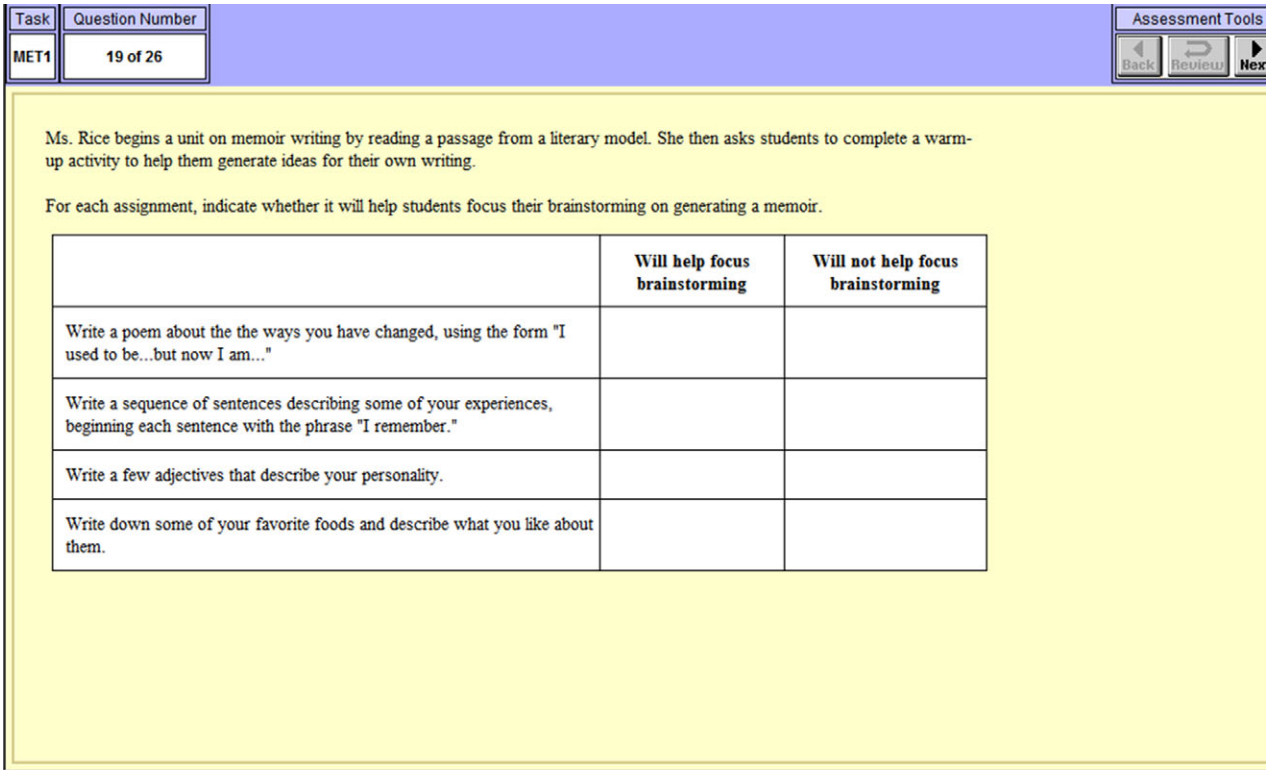



**Figure B19** Question #19.

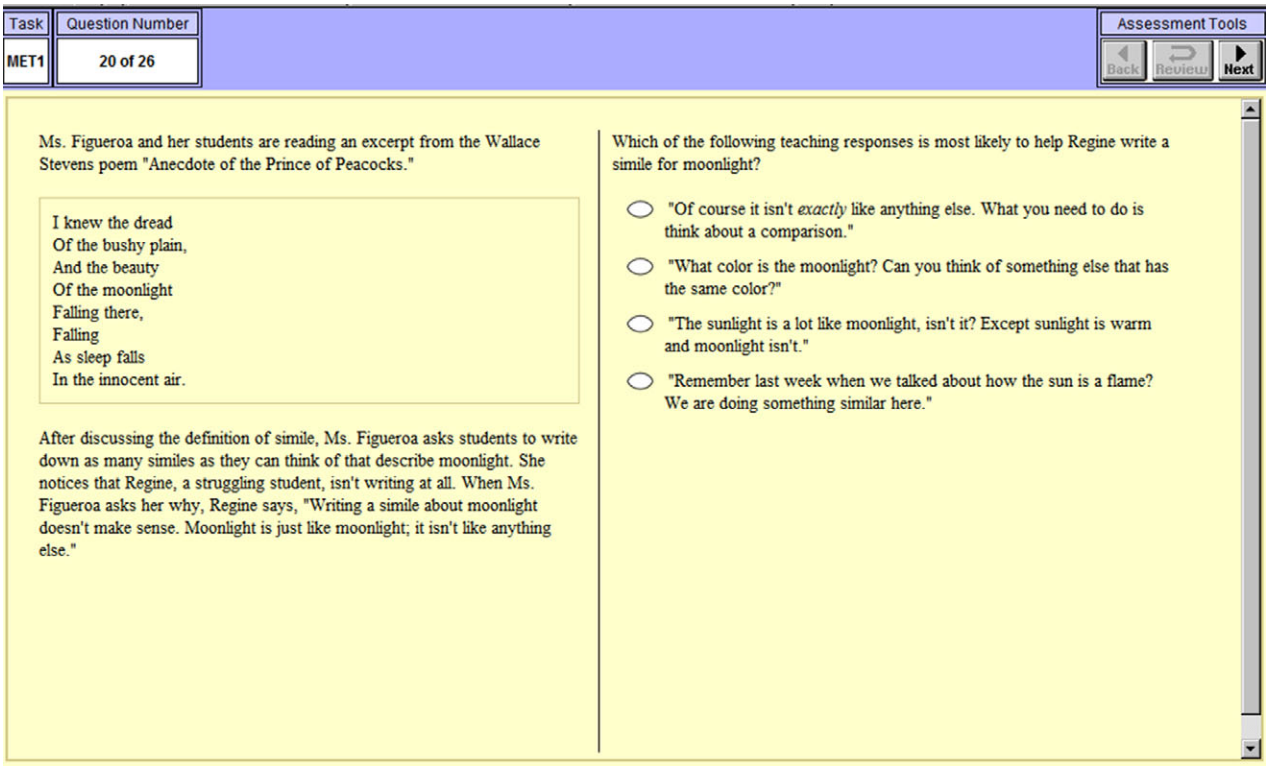



**Figure B20** Question #20.

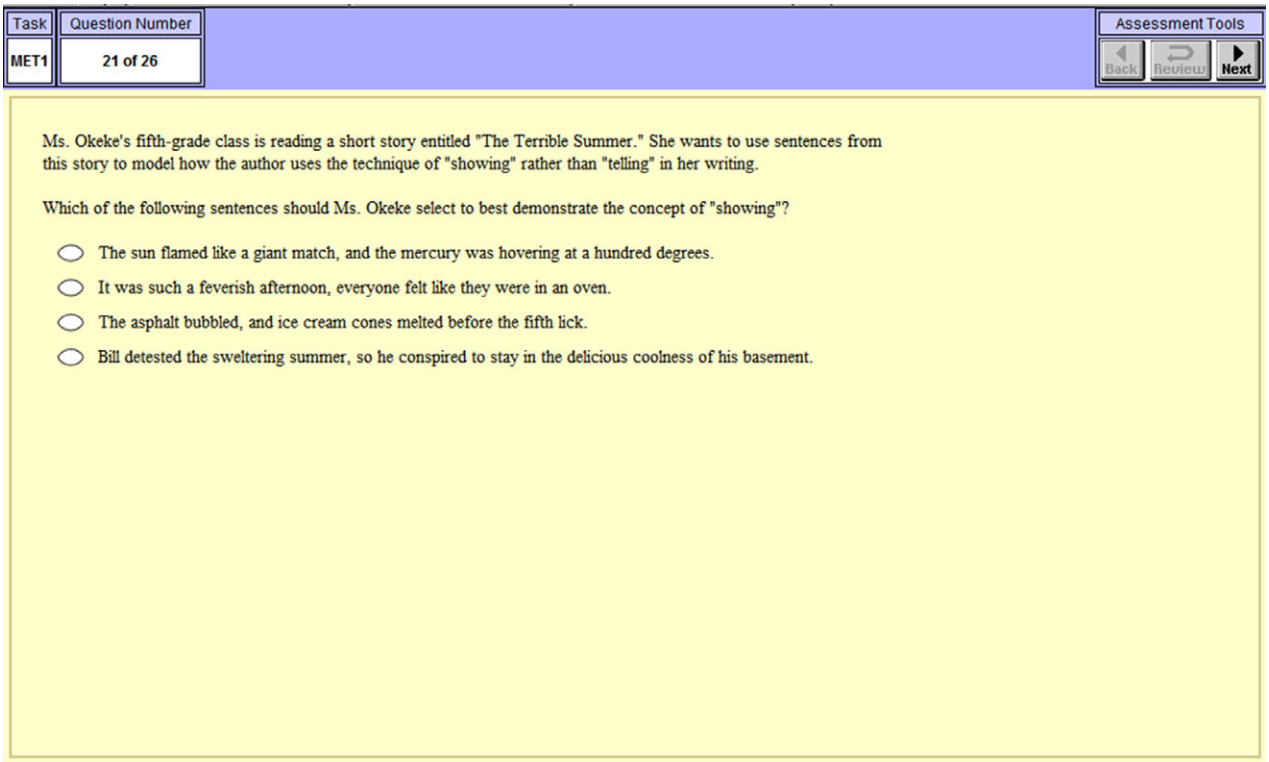



**Figure B21** Question #21.

**Figure B22** Question #22. (Question #22 in this assessment cannot be shown due to copyright restrictions for an excerpt used from a published text in this question.)

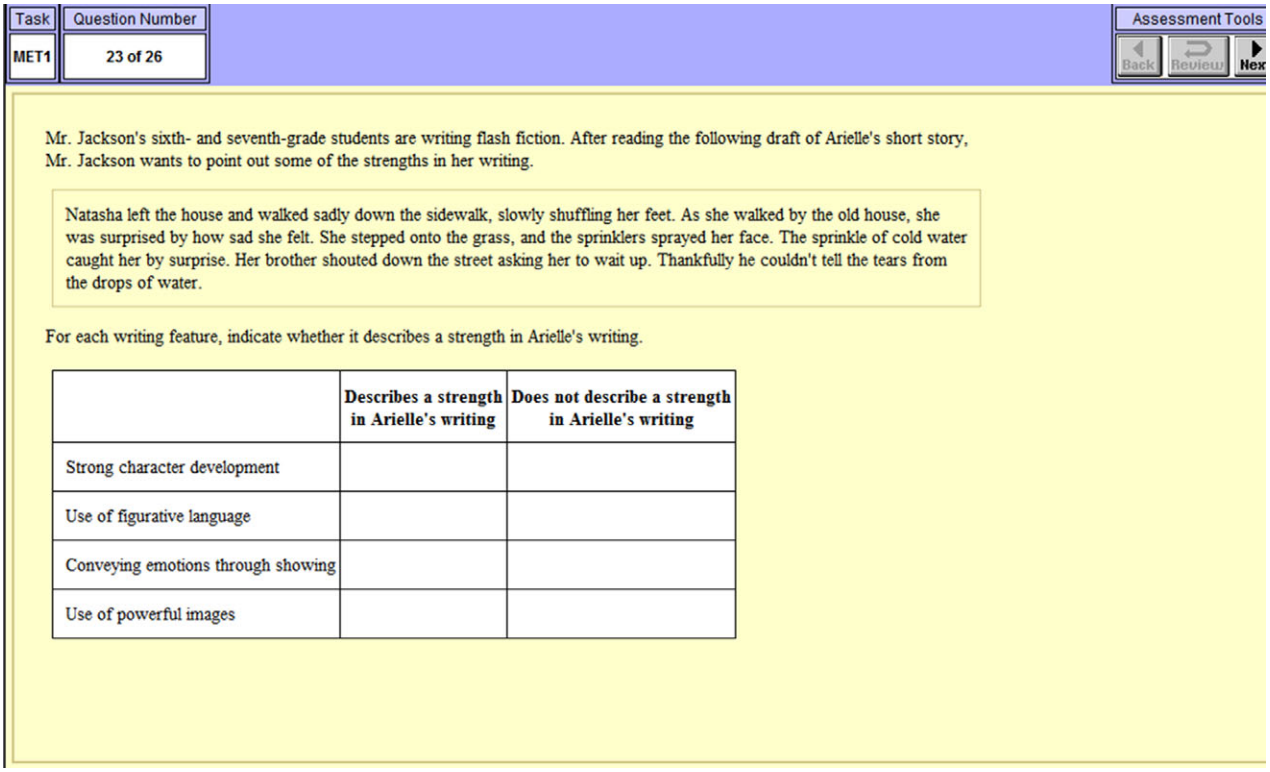



**Figure B23** Question #23.

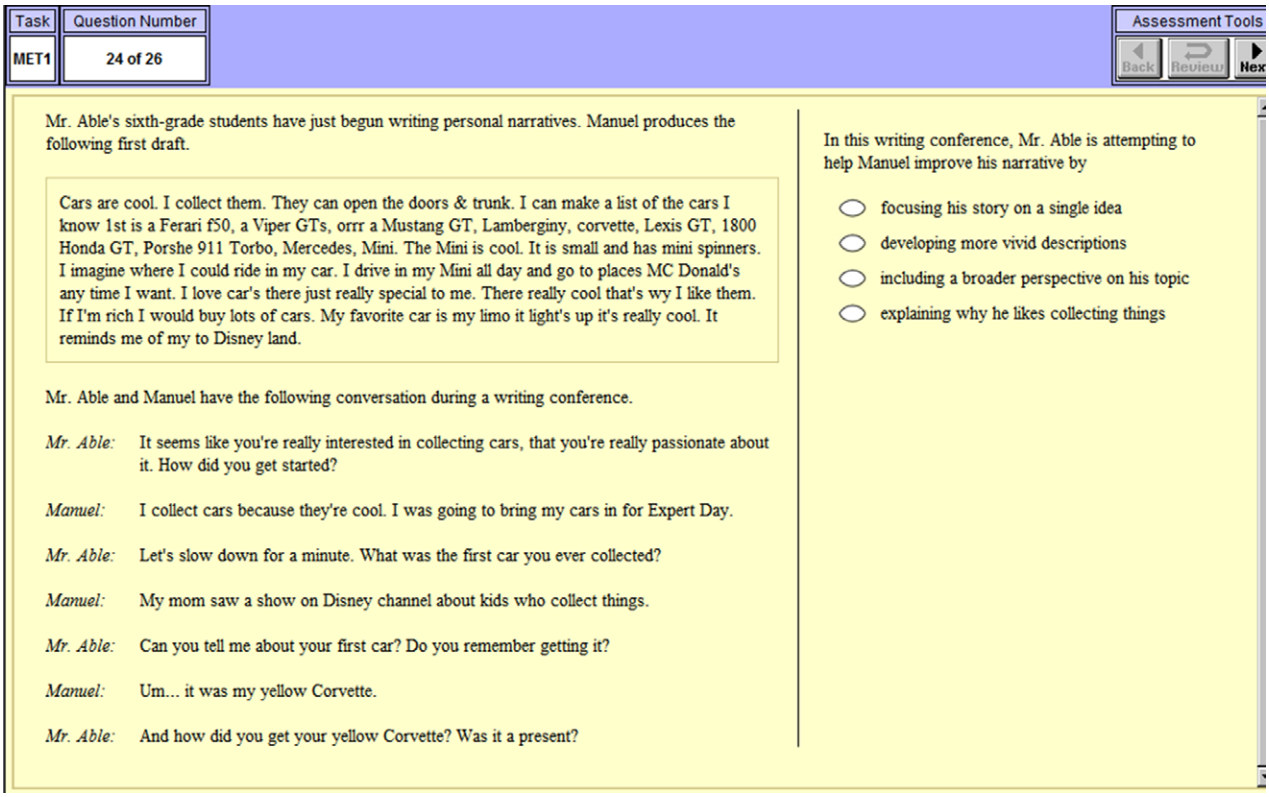



**Figure B24** Question #24.

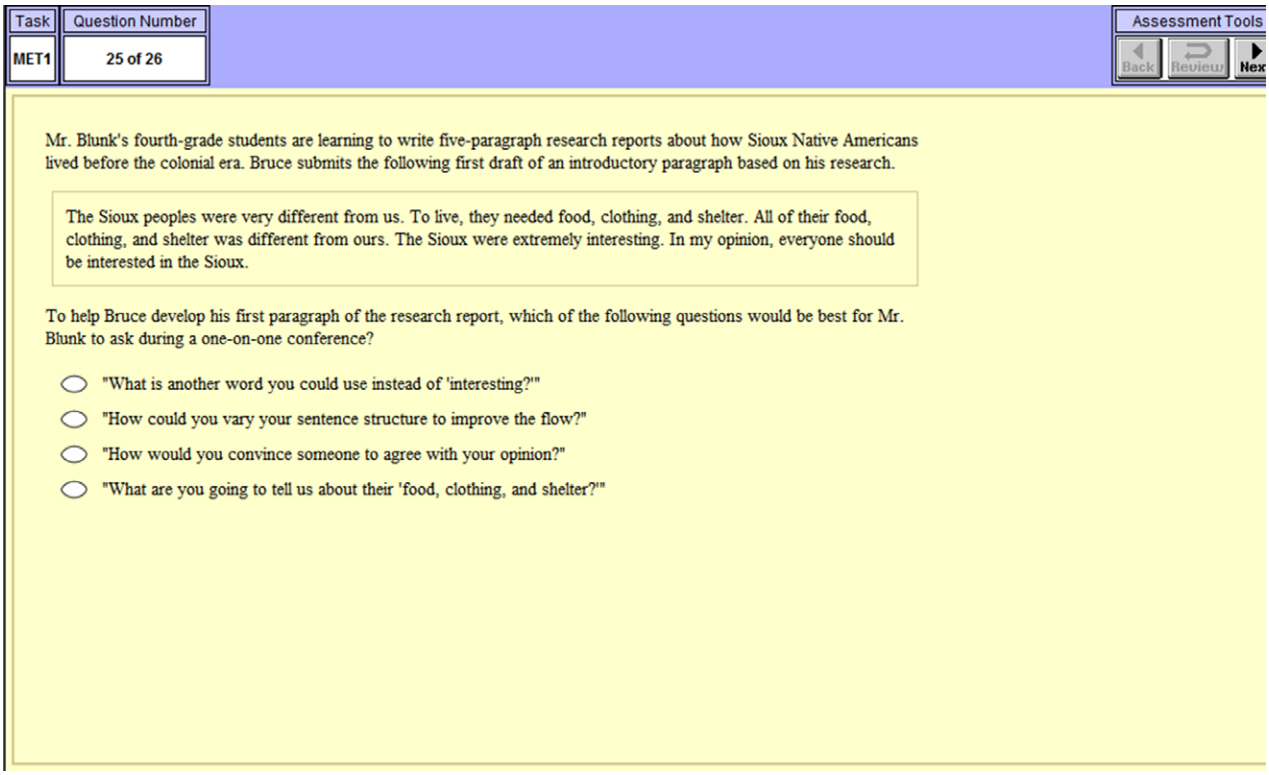



**Figure B25** Question #25.

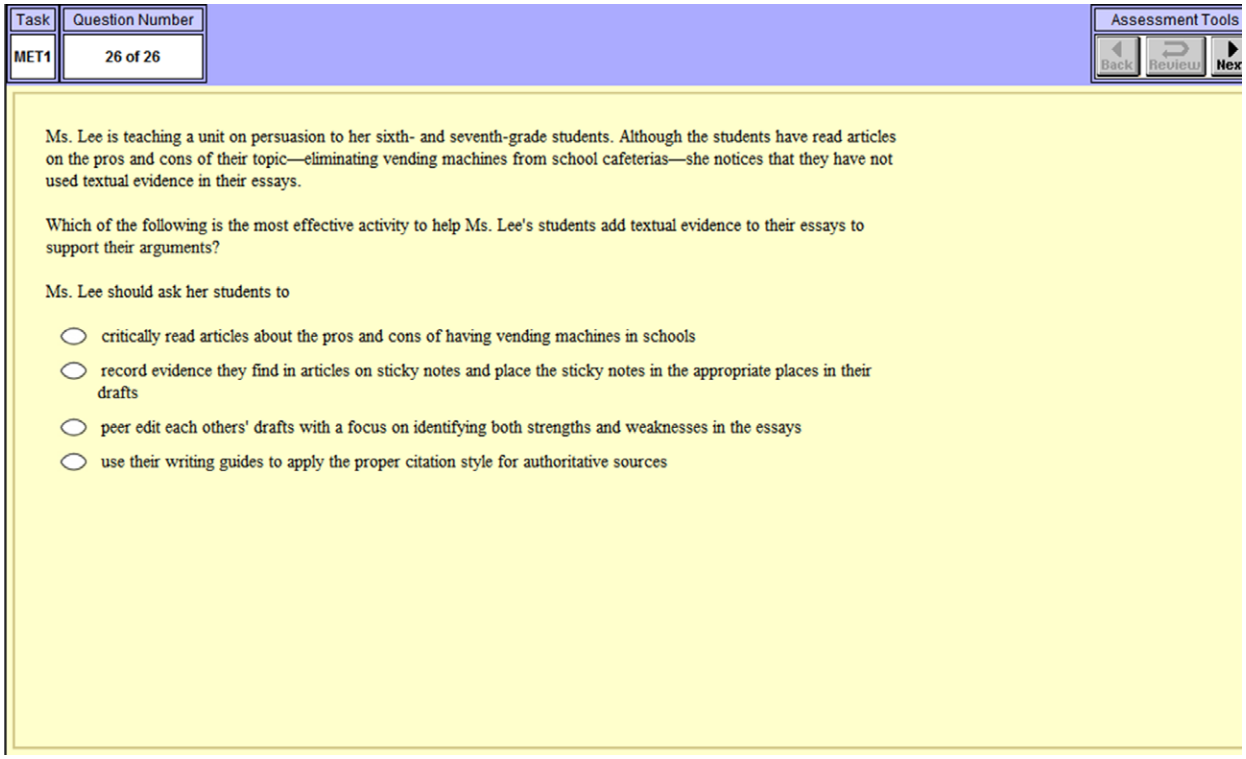



**Figure B26** Question #26.

## Appendix C

## Mathematics 6–8 Assessment

**Table C1** Item/Question Sequence Number for Mathematics 6–8 Assessment

| Item/question sequence number | Answer key | Item/question sequence number | Answer key |
|---|---|---|---|
| 1 | C | 11e | Provides |
| 2 | C | 11f | Does not provide |
| 3a | Provides | 12 | B |
| 3b | Provides | 13 | A |
| 3c | Does not provide | 14a | Provides |
| 4 | B | 14b | Does not provide |
| 5 | B | 14c | Provides |
| 6a | Provides | 14d | Provides |
| 6b | Does not provide | 15 | B |
| 6c | Provides | 16a | Provides |
| 6d | Provides | 16b | Provides |
| 6e | Provides | 16c | Provides |
| 7 | A | 16d | Does not provide |
| 8a | Valid | 16e | Does not provide |
| 8b | Valid | 17 | A |
| 8c | Not valid | 18 | B |
| 8d | Valid | 19a | Does not demonstrate |
| 8e | Not valid | 19b | Demonstrates |
| 9 | D | 19c | Demonstrates |
| 10 | B | 19d | Demonstrates |
| 11a | Provides | 20 | B |
| 11b | Provides | 21 | C |
| 11c | Does not provide | 22 | Open-ended response |
| 11d | Provides | 23 | Open-ended response |


**Figure C1** Question #1.

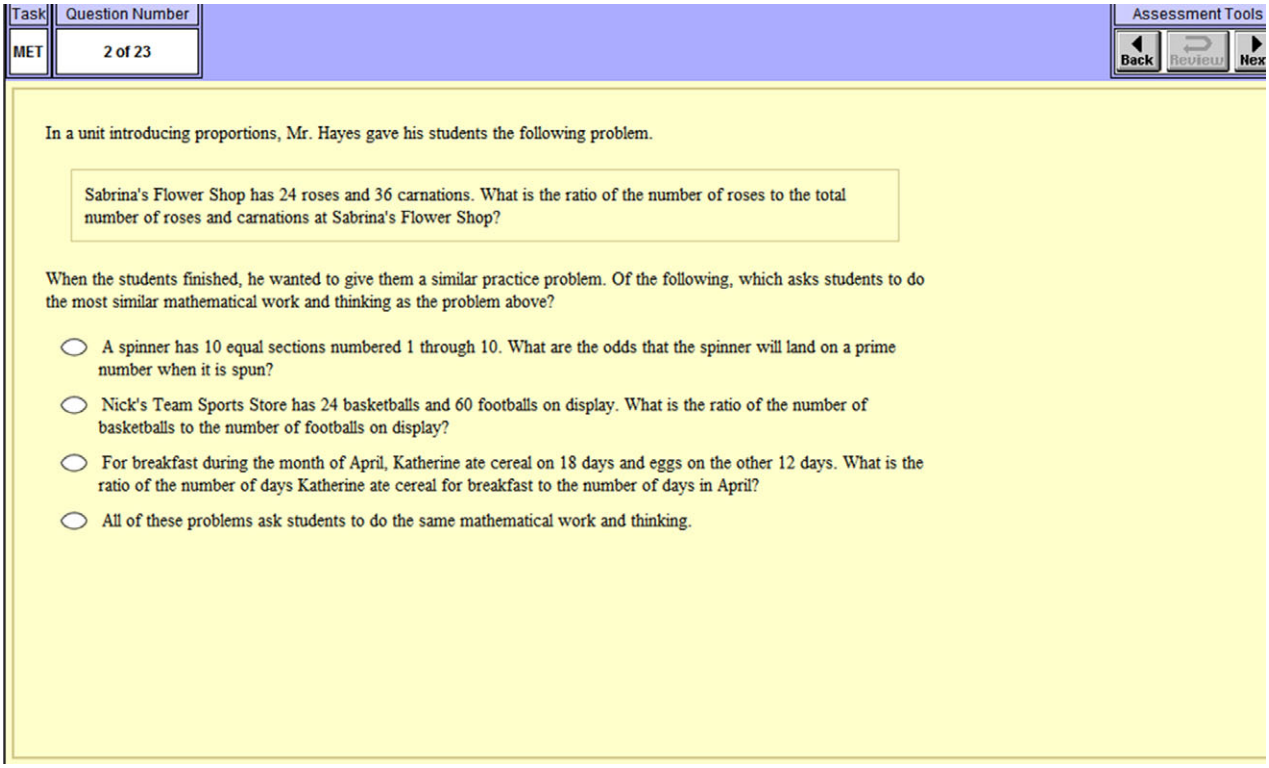
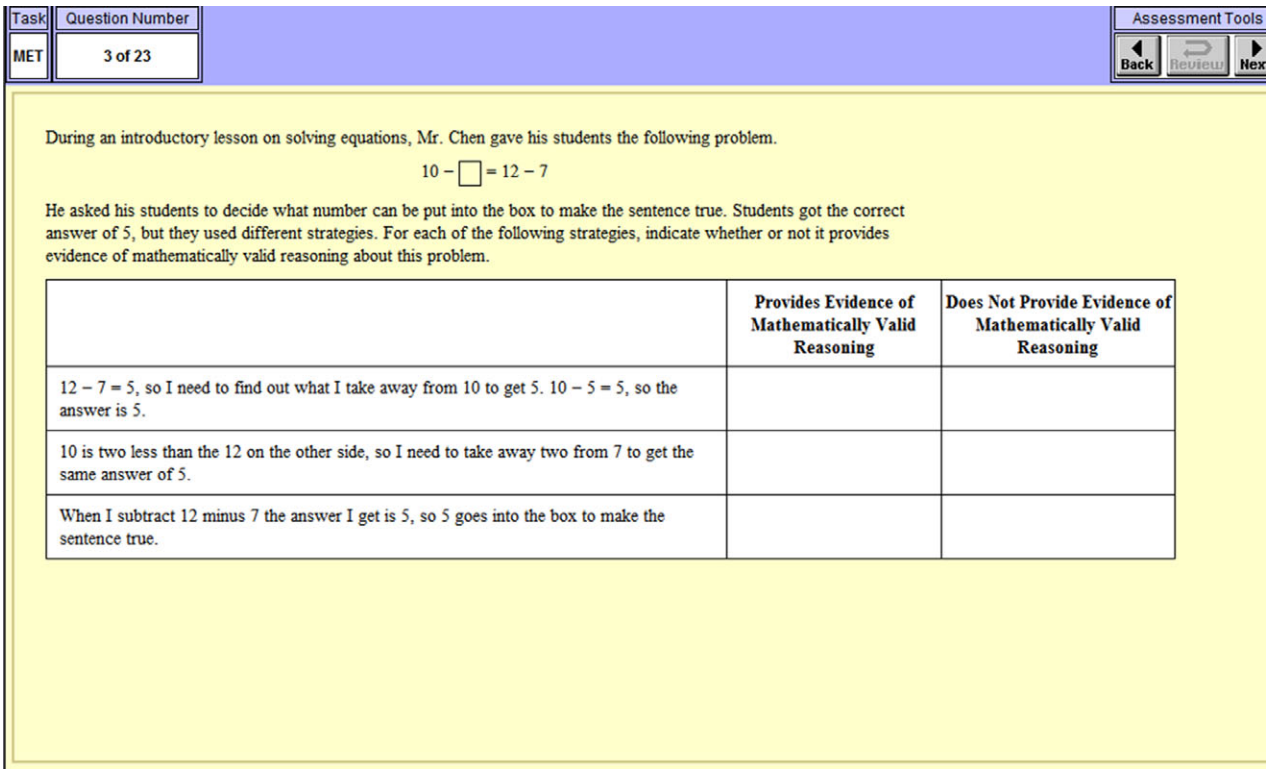



**Figure C2** Question #2.



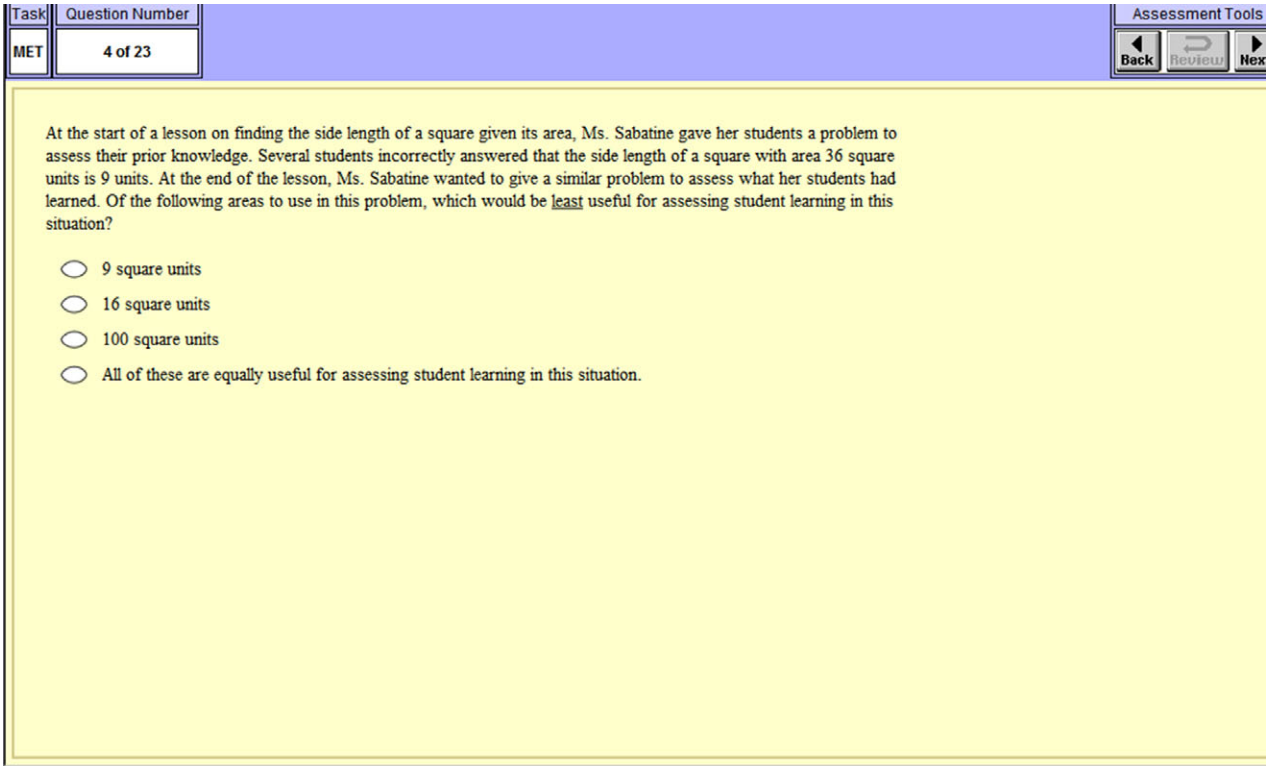

During an introductory lesson on solving equations, Mr. Chen gave his students the following problem.

$$10 - \square = 12 - 7$$

He asked his students to decide what number can be put into the box to make the sentence true. Students got the correct answer of 5, but they used different strategies. For each of the following strategies, indicate whether or not it provides evidence of mathematically valid reasoning about this problem.

| | Provides Evidence of Mathematically Valid Reasoning | Does Not Provide Evidence of Mathematically Valid Reasoning |
|---|---|---|
| 12 − 7 = 5, so I need to find out what I take away from 10 to get 5. 10 − 5 = 5, so the answer is 5. | | |
| 10 is two less than the 12 on the other side, so I need to take away two from 7 to get the same answer of 5. | | |
| When I subtract 12 minus 7 the answer I get is 5, so 5 goes into the box to make the sentence true. | | |

**Figure C3** Question #3.




At the start of a lesson on finding the side length of a square given its area, Ms. Sabatine gave her students a problem to assess their prior knowledge. Several students incorrectly answered that the side length of a square with area 36 square units is 9 units. At the end of the lesson, Ms. Sabatine wanted to give a similar problem to assess what her students had learned. Of the following areas to use in this problem, which would be least useful for assessing student learning in this situation?

- ◯ 9 square units
- ◯ 16 square units
- ◯ 100 square units
- ◯ All of these are equally useful for assessing student learning in this situation.

**Figure C4** Question #4.

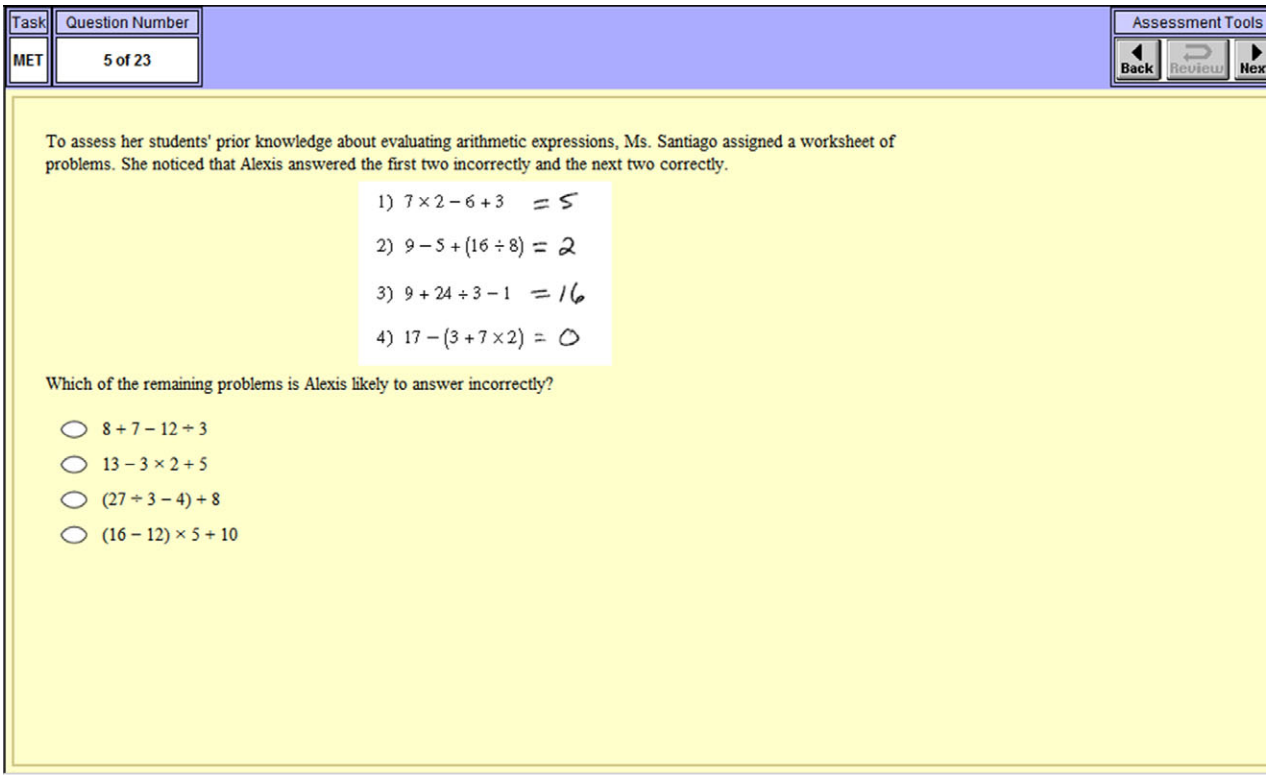
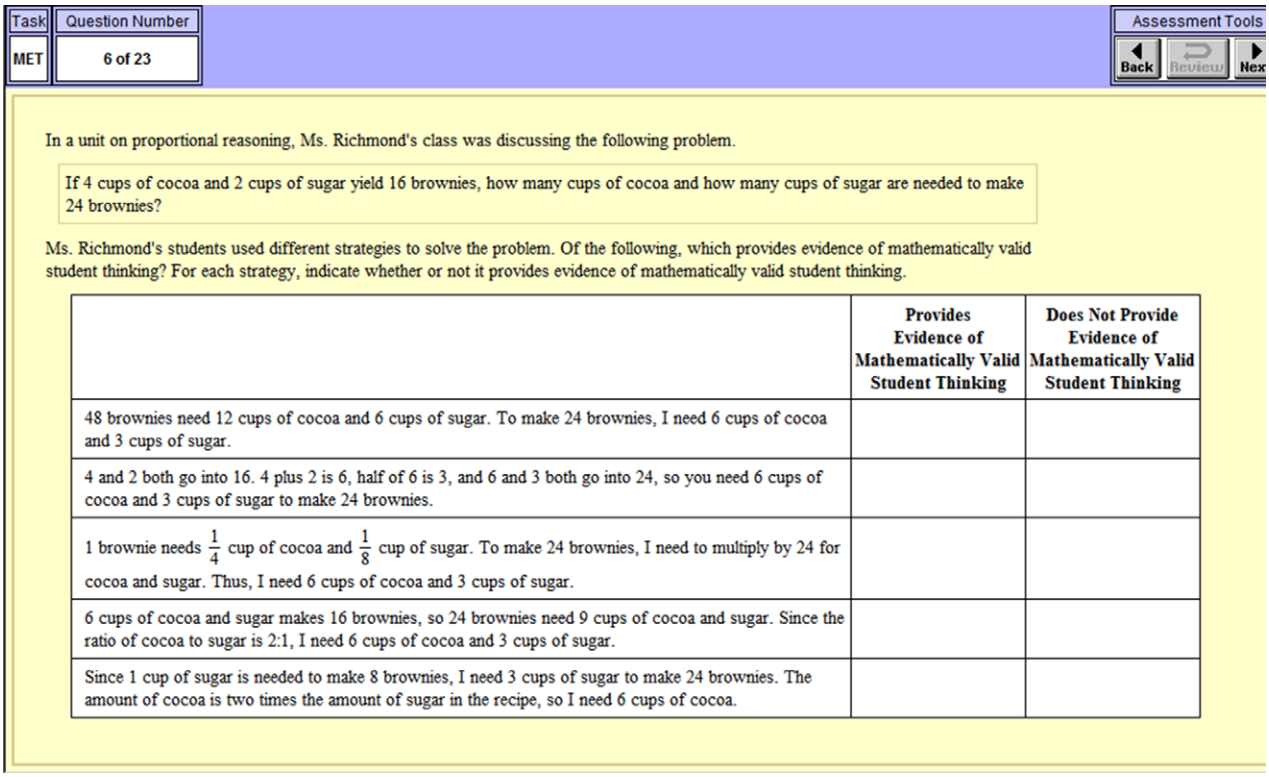



**Figure C5** Question #5.




| | Provides Evidence of Mathematically Valid Student Thinking | Does Not Provide Evidence of Mathematically Valid Student Thinking |
|---|---|---|
| 48 brownies need 12 cups of cocoa and 6 cups of sugar. To make 24 brownies, I need 6 cups of cocoa and 3 cups of sugar. | | |
| 4 and 2 both go into 16. 4 plus 2 is 6, half of 6 is 3, and 6 and 3 both go into 24, so you need 6 cups of cocoa and 3 cups of sugar to make 24 brownies. | | |
| 1 brownie needs $\frac{1}{4}$ cup of cocoa and $\frac{1}{8}$ cup of sugar. To make 24 brownies, I need to multiply by 24 for cocoa and sugar. Thus, I need 6 cups of cocoa and 3 cups of sugar. | | |
| 6 cups of cocoa and sugar makes 16 brownies, so 24 brownies need 9 cups of cocoa and sugar. Since the ratio of cocoa to sugar is 2:1, I need 6 cups of cocoa and 3 cups of sugar. | | |
| Since 1 cup of sugar is needed to make 8 brownies, I need 3 cups of sugar to make 24 brownies. The amount of cocoa is two times the amount of sugar in the recipe, so I need 6 cups of cocoa. | | |

**Figure C6** Question #6.

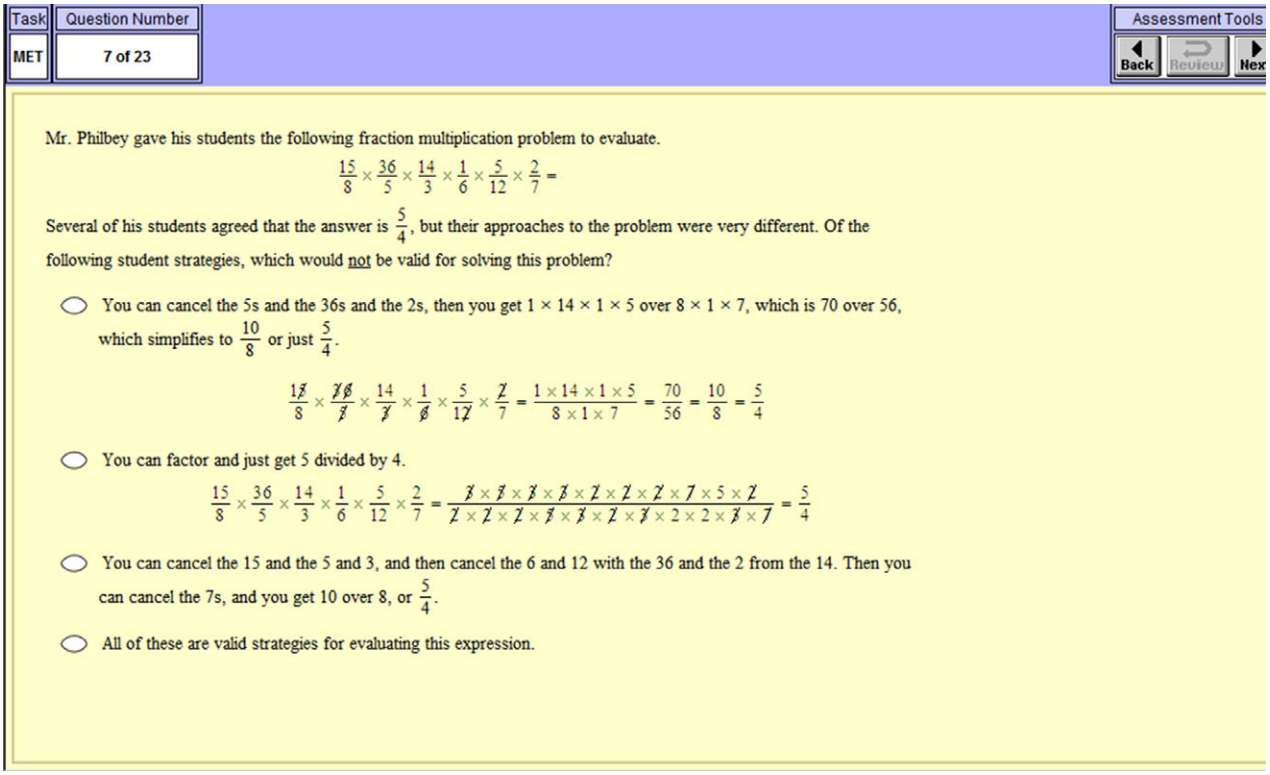



**Figure C7** Question #7.

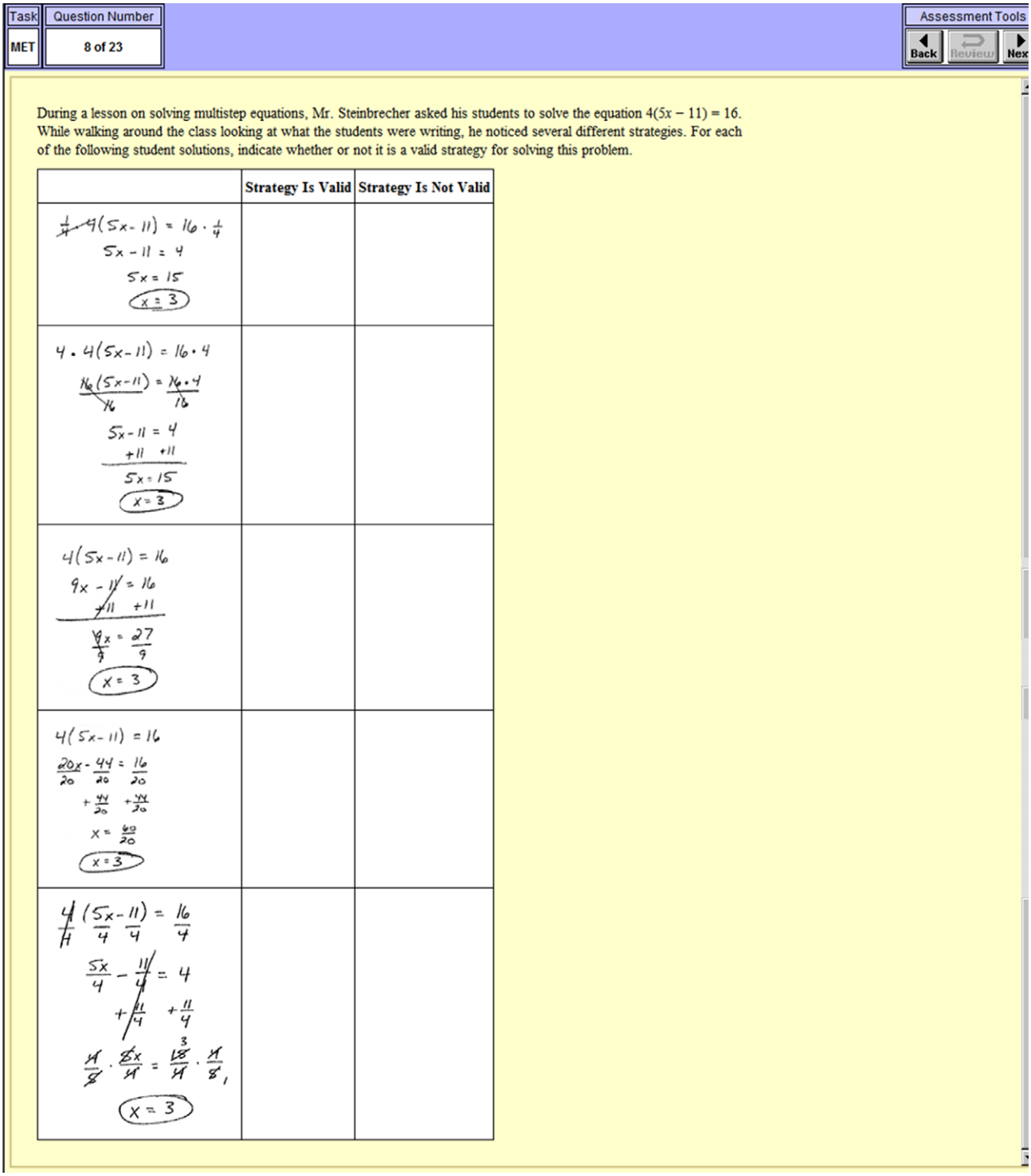



**Figure C8** Question #8.


**Figure C9** Question #9.

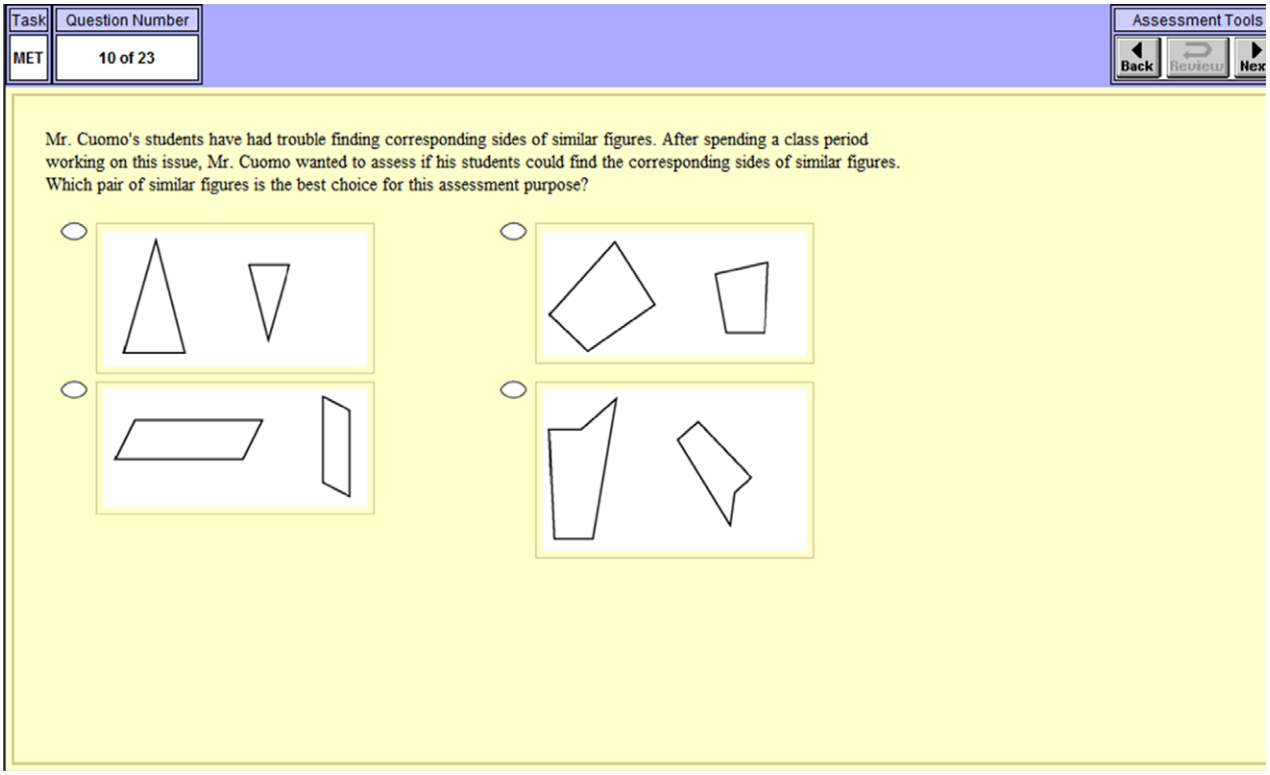
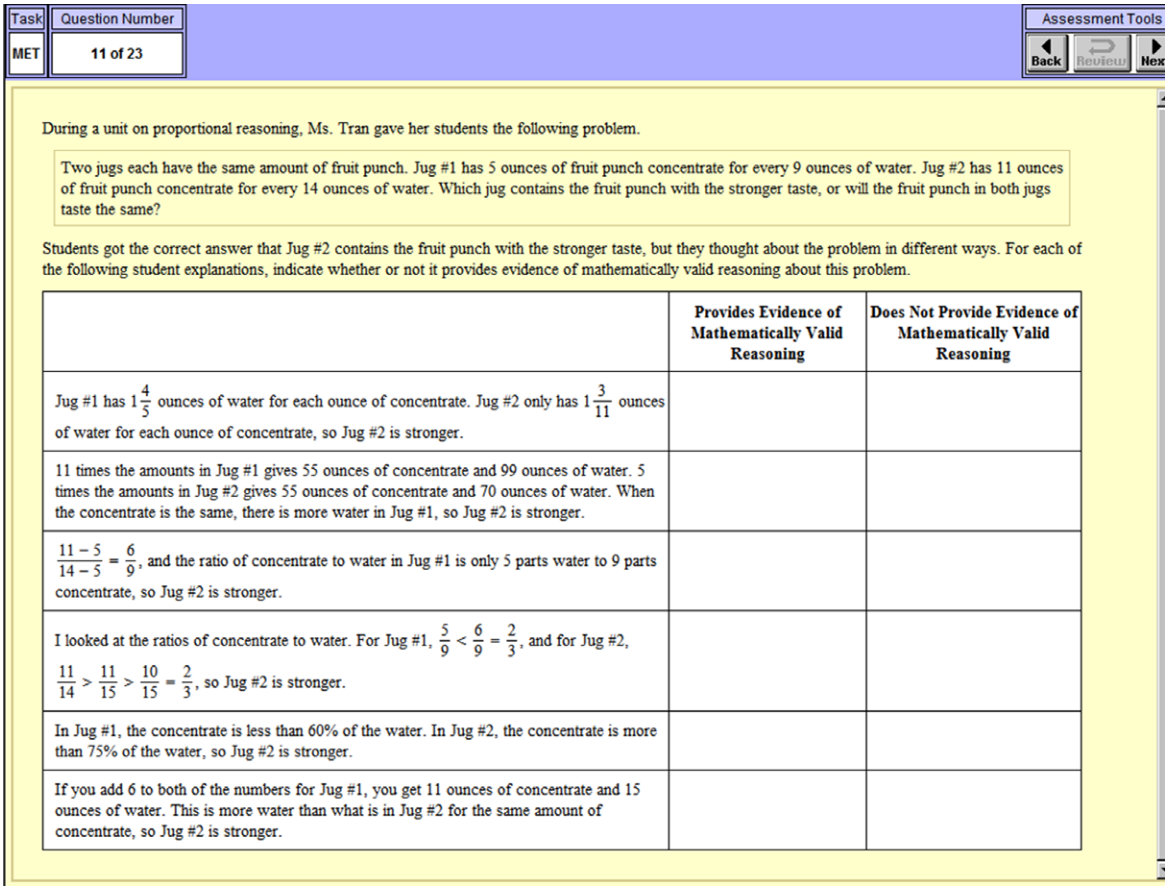



**Figure C10** Question #10.



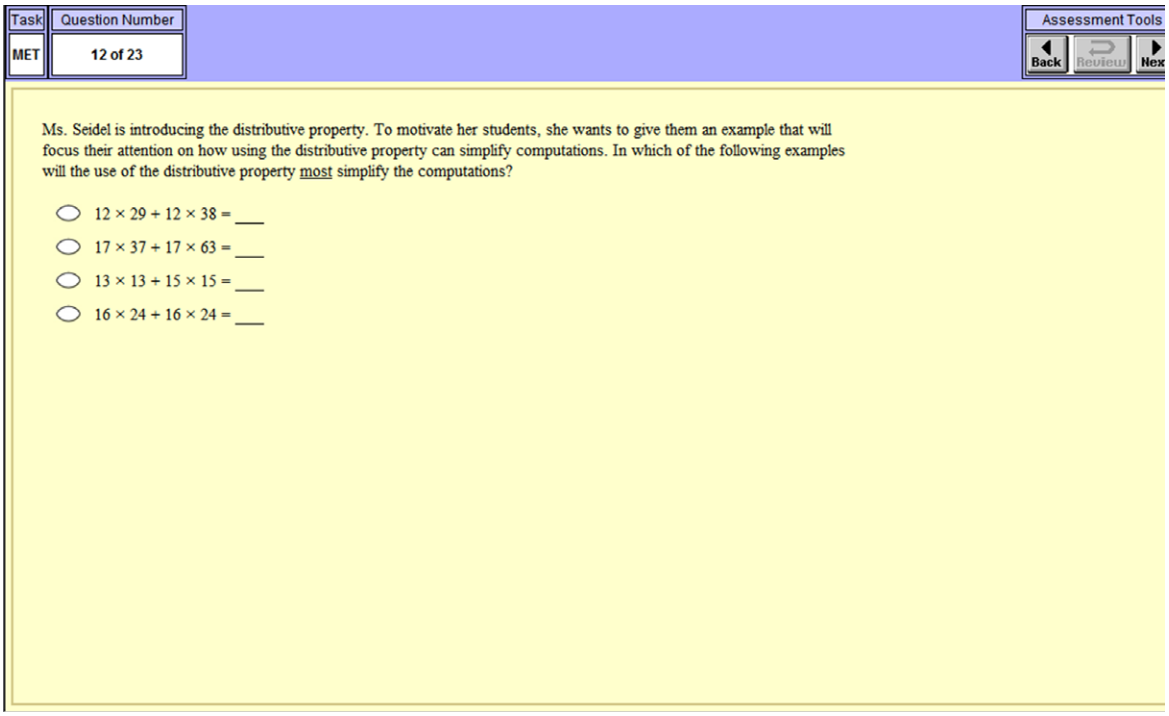

During a unit on proportional reasoning, Ms. Tran gave her students the following problem.

> Two jugs each have the same amount of fruit punch. Jug #1 has 5 ounces of fruit punch concentrate for every 9 ounces of water. Jug #2 has 11 ounces of fruit punch concentrate for every 14 ounces of water. Which jug contains the fruit punch with the stronger taste, or will the fruit punch in both jugs taste the same?

Students got the correct answer that Jug #2 contains the fruit punch with the stronger taste, but they thought about the problem in different ways. For each of the following student explanations, indicate whether or not it provides evidence of mathematically valid reasoning about this problem.

| | **Provides Evidence of Mathematically Valid Reasoning** | **Does Not Provide Evidence of Mathematically Valid Reasoning** |
|---|---|---|
| Jug #1 has $1\frac{4}{5}$ ounces of water for each ounce of concentrate. Jug #2 only has $1\frac{3}{11}$ ounces of water for each ounce of concentrate, so Jug #2 is stronger. | | |
| 11 times the amounts in Jug #1 gives 55 ounces of concentrate and 99 ounces of water. 5 times the amounts in Jug #2 gives 55 ounces of concentrate and 70 ounces of water. When the concentrate is the same, there is more water in Jug #1, so Jug #2 is stronger. | | |
| $\frac{11-5}{14-5} = \frac{6}{9}$, and the ratio of concentrate to water in Jug #1 is only 5 parts water to 9 parts concentrate, so Jug #2 is stronger. | | |
| I looked at the ratios of concentrate to water. For Jug #1, $\frac{5}{9} < \frac{6}{9} = \frac{2}{3}$, and for Jug #2, $\frac{11}{14} > \frac{11}{15} > \frac{10}{15} = \frac{2}{3}$, so Jug #2 is stronger. | | |
| In Jug #1, the concentrate is less than 60% of the water. In Jug #2, the concentrate is more than 75% of the water, so Jug #2 is stronger. | | |
| If you add 6 to both of the numbers for Jug #1, you get 11 ounces of concentrate and 15 ounces of water. This is more water than what is in Jug #2 for the same amount of concentrate, so Jug #2 is stronger. | | |

**Figure C11** Question #11.




Ms. Seidel is introducing the distributive property. To motivate her students, she wants to give them an example that will focus their attention on how using the distributive property can simplify computations. In which of the following examples will the use of the distributive property most simplify the computations?

- ○ $12 \times 29 + 12 \times 38 =$ ___
- ○ $17 \times 37 + 17 \times 63 =$ ___
- ○ $13 \times 13 + 15 \times 15 =$ ___
- ○ $16 \times 24 + 16 \times 24 =$ ___

**Figure C12** Question #12.

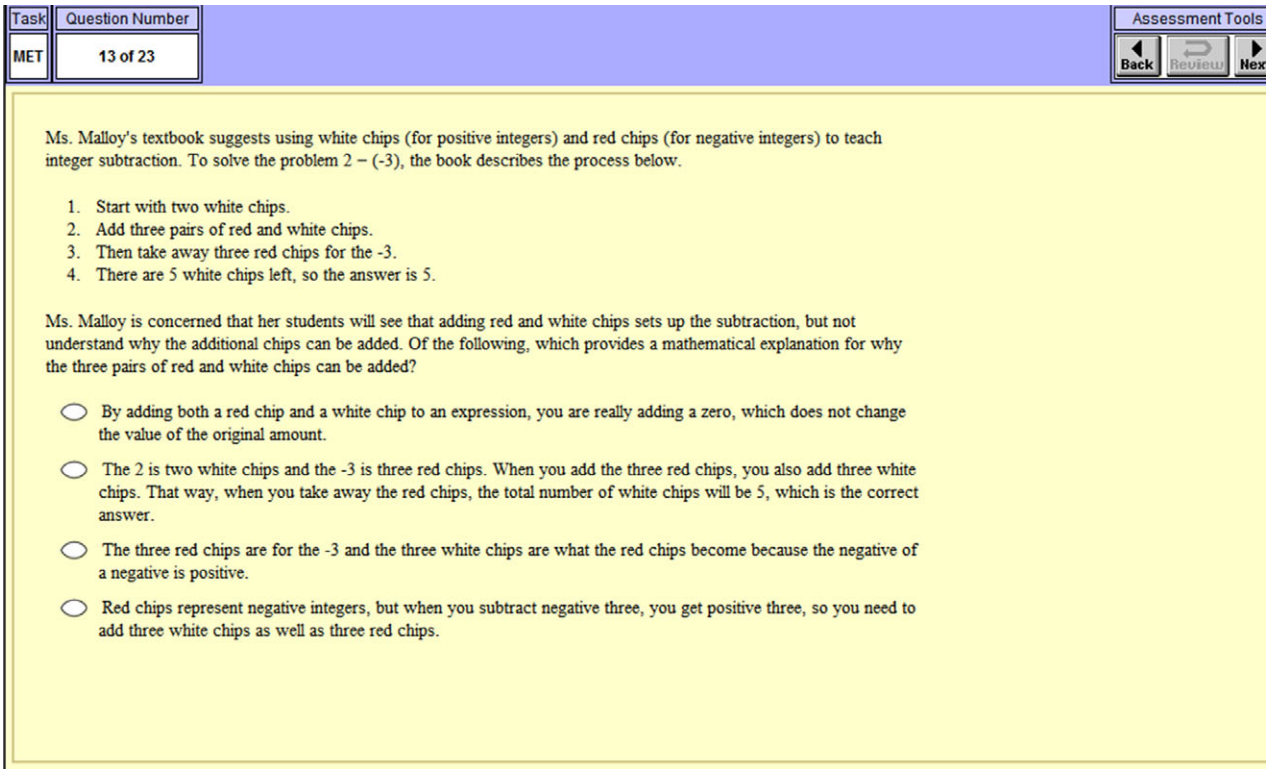



**Figure C13** Question #13.

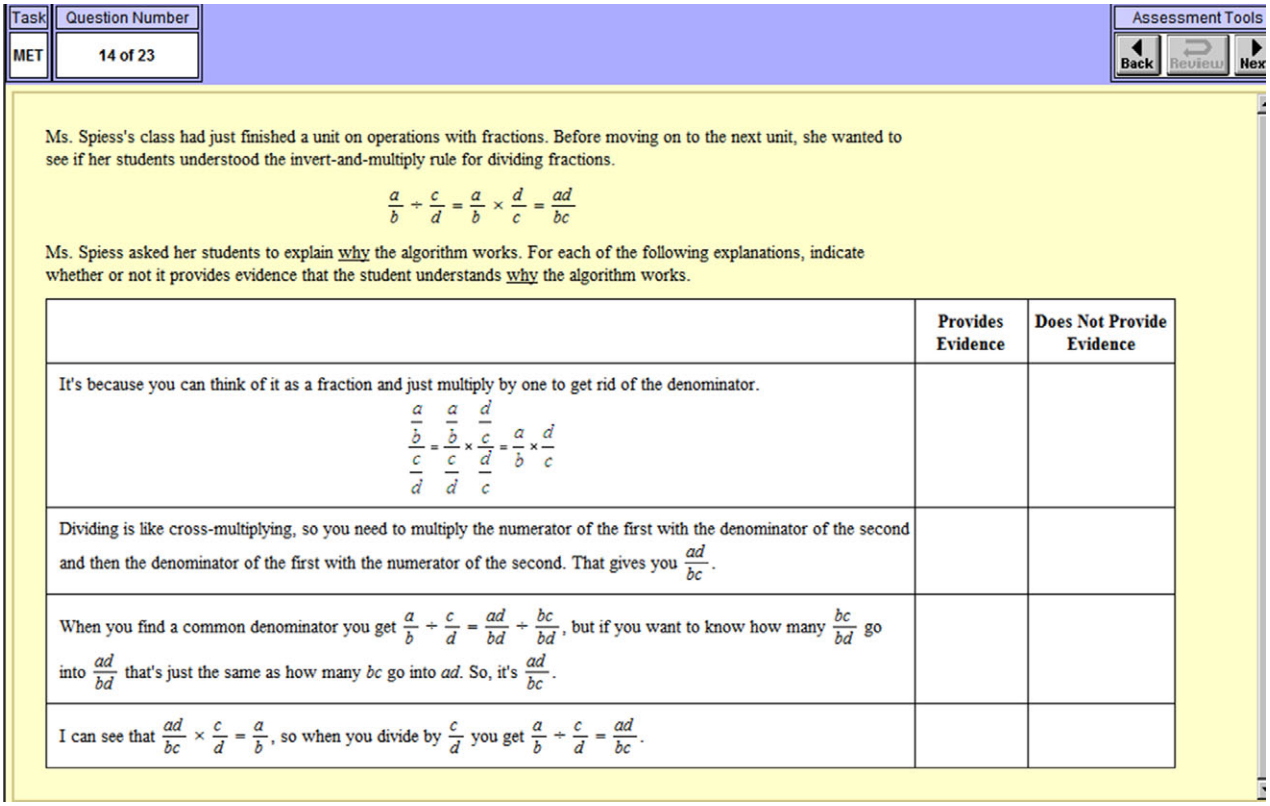



**Figure C14** Question #14.

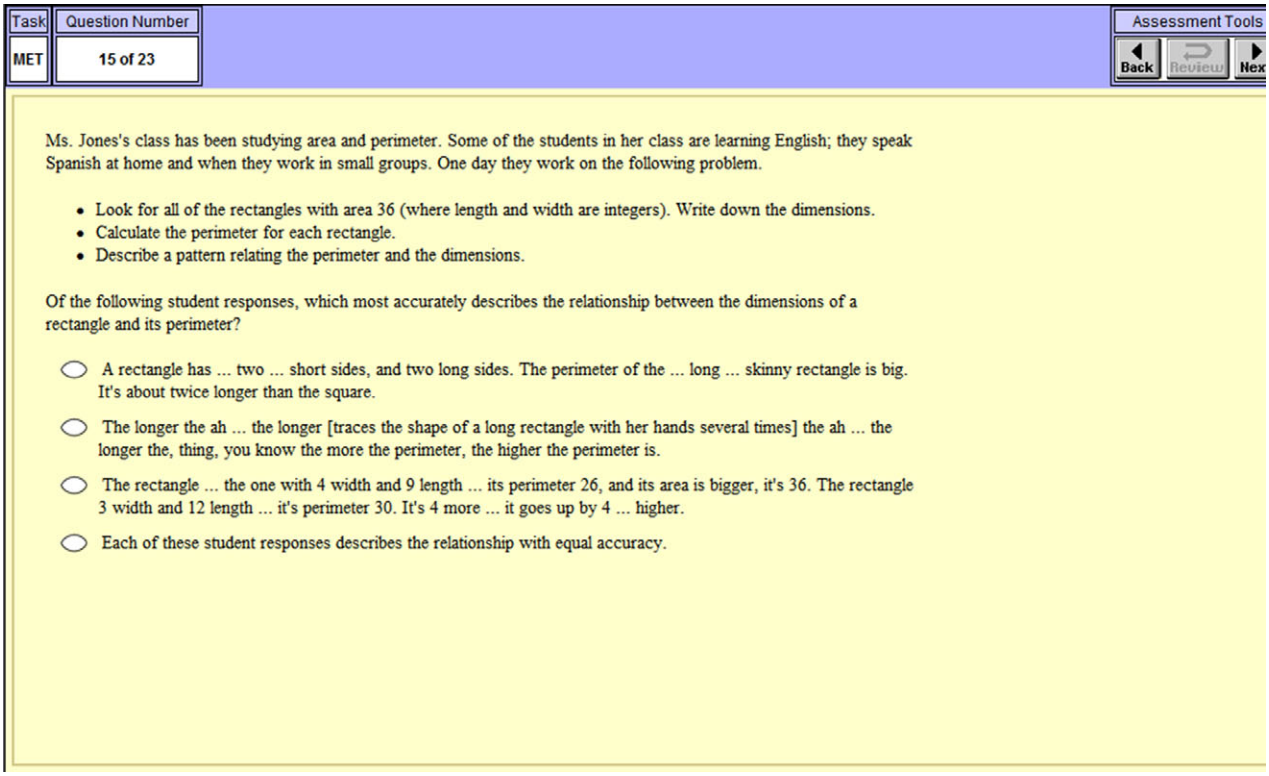



**Figure C15** Question #15.

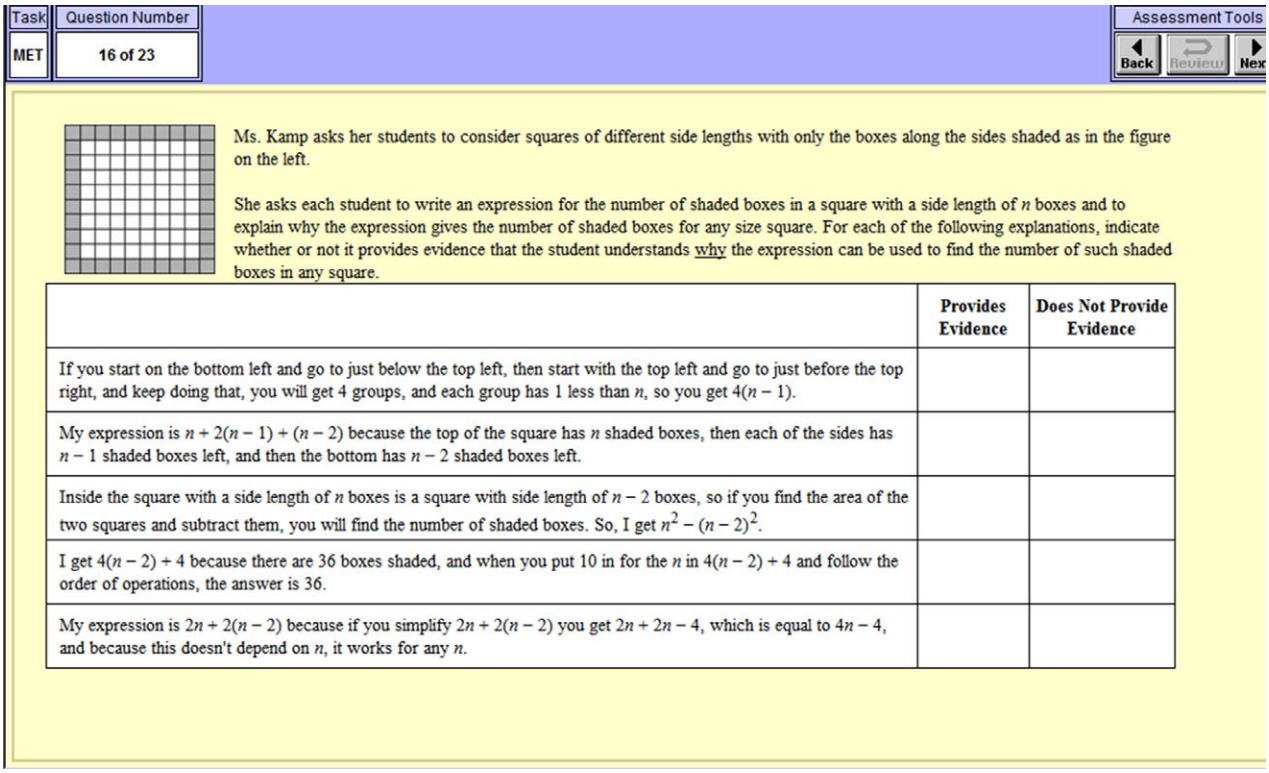



**Figure C16** Question #16.

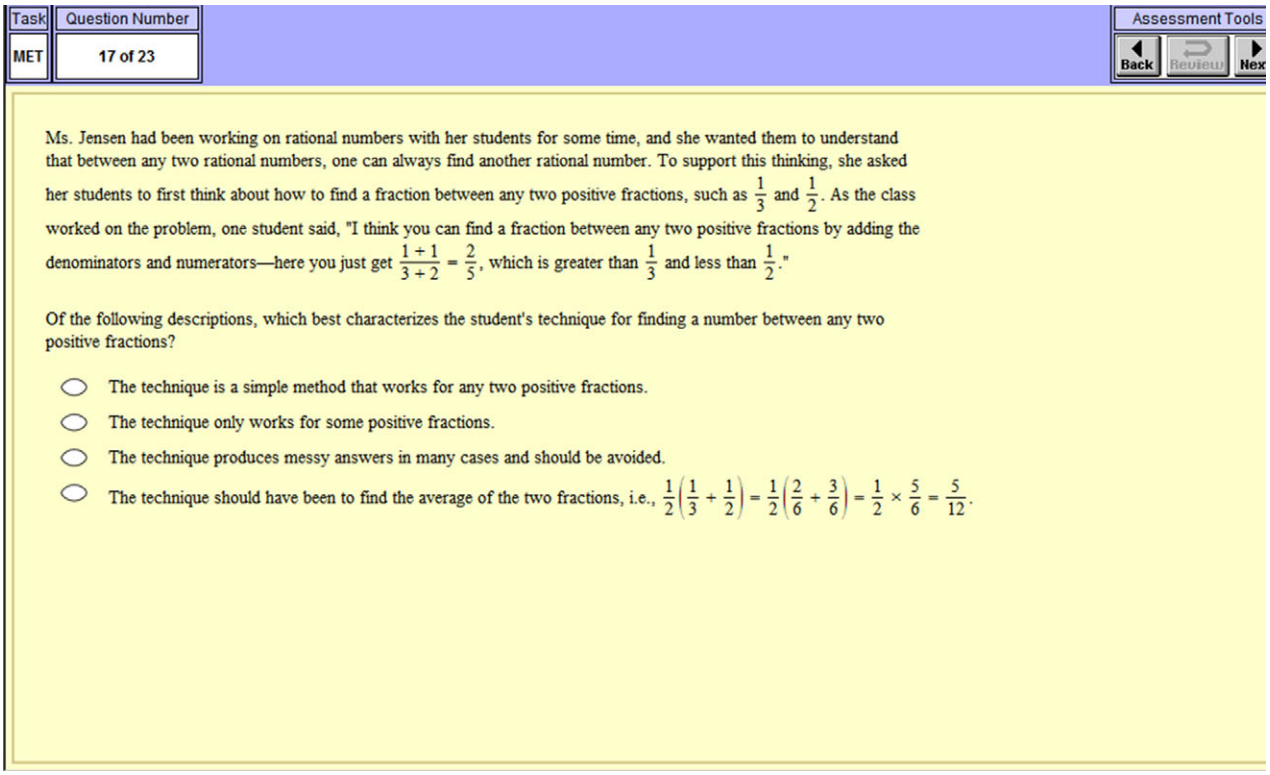



**Figure C17** Question #17.

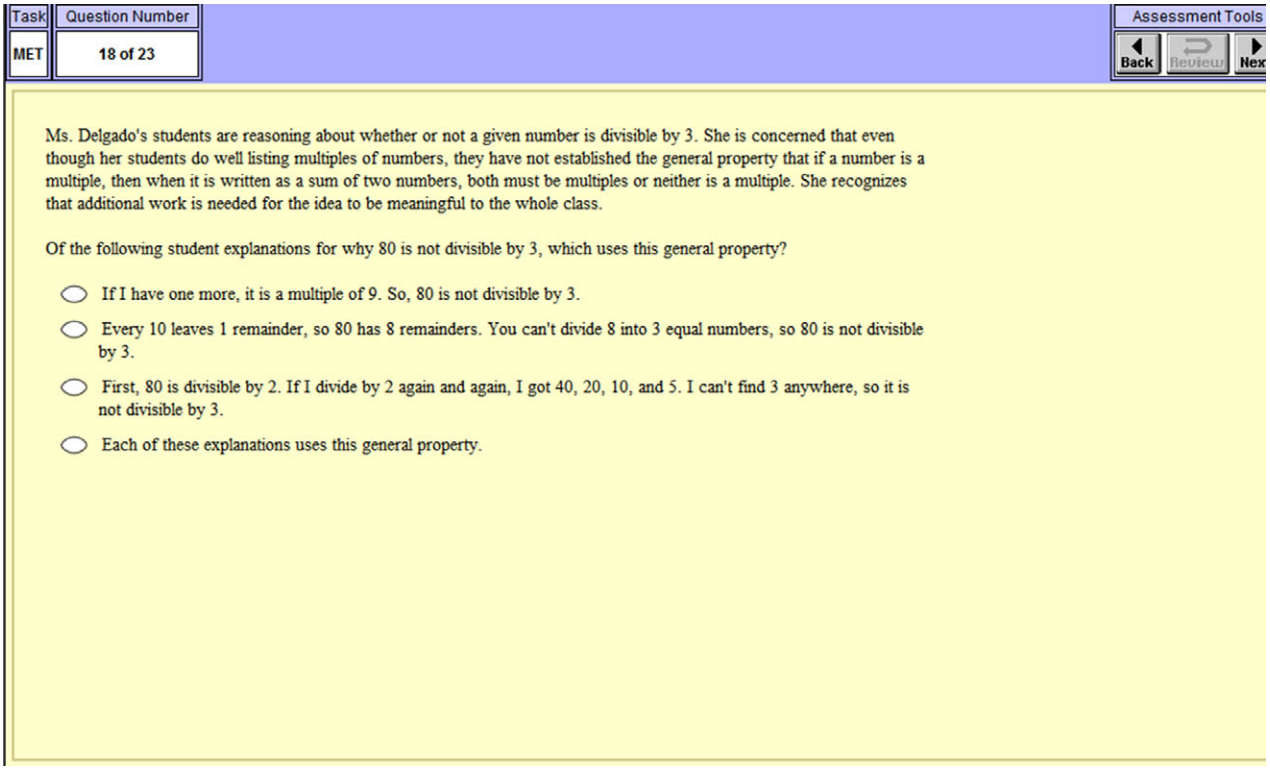



**Figure C18** Question #18.

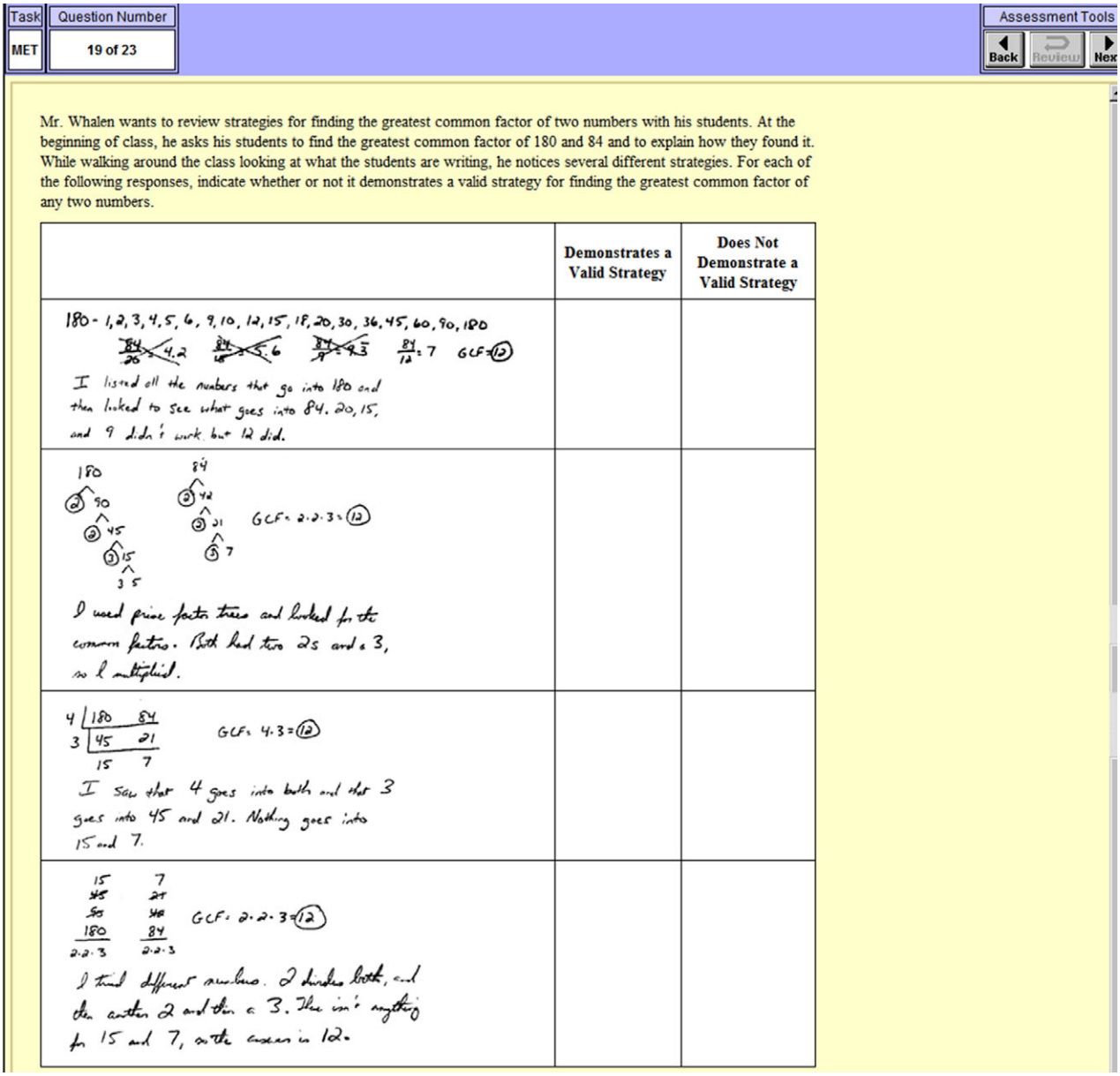



**Figure C19** Question #19.

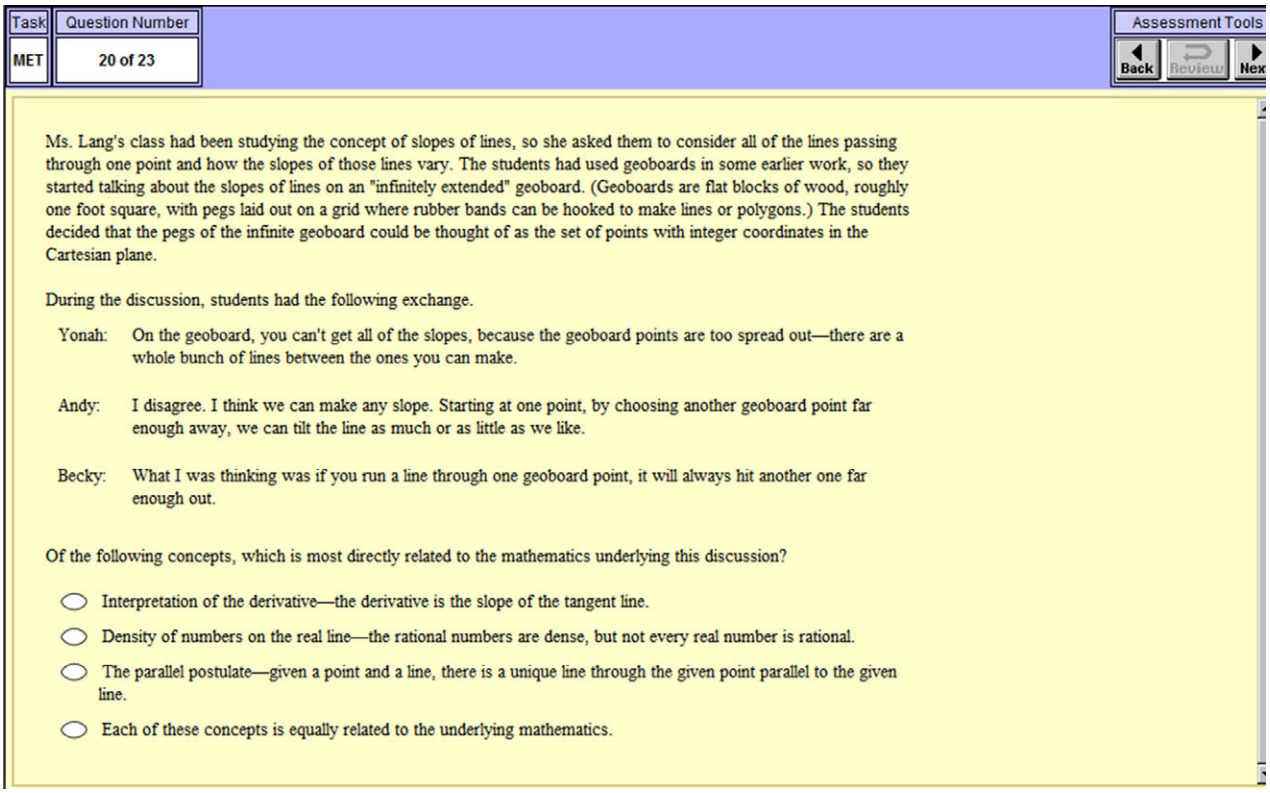



**Figure C20** Question #20.

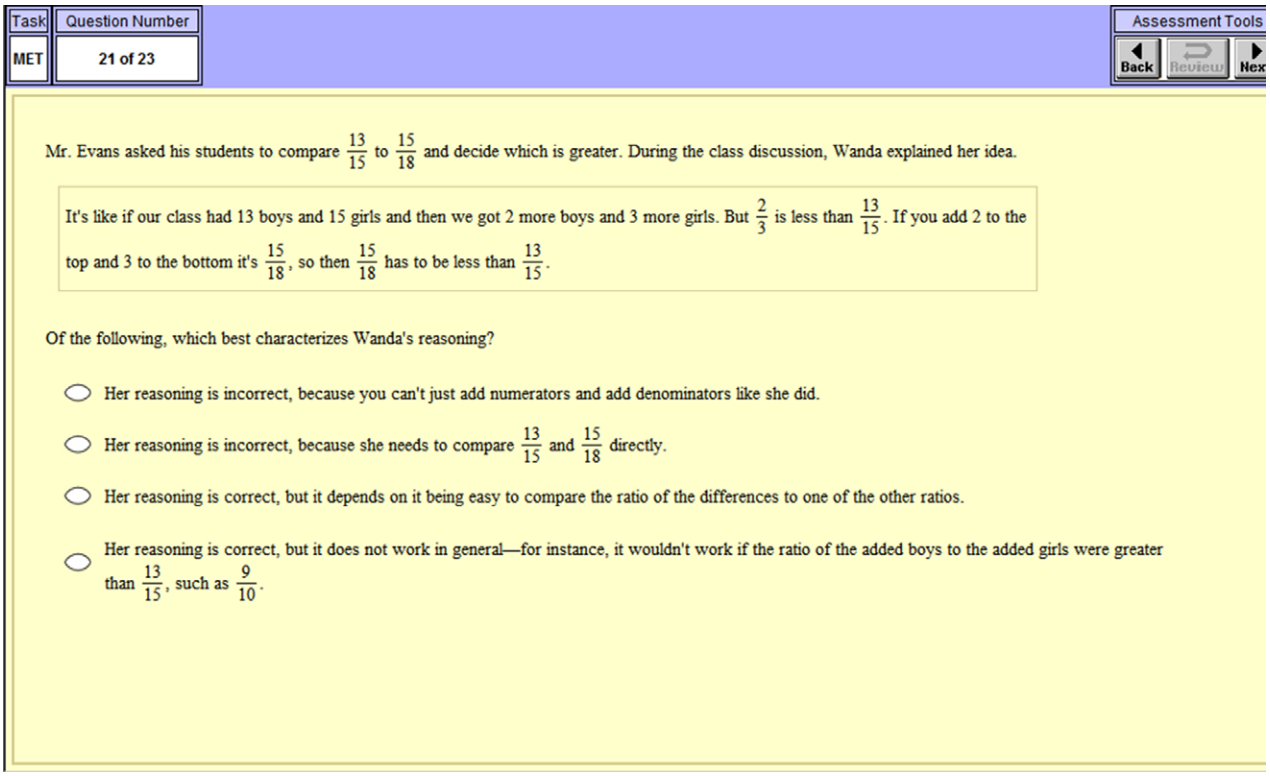



**Figure C21** Question #21.

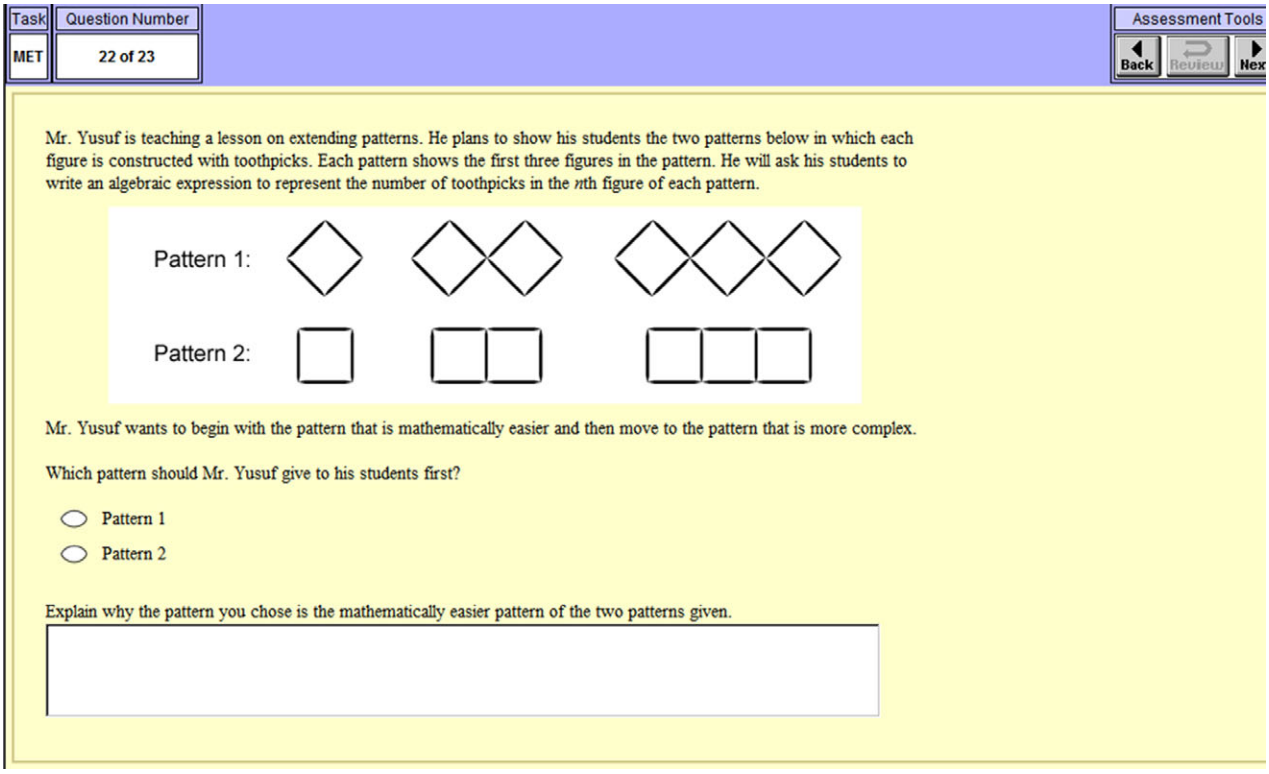



**Figure C22** Question #22.

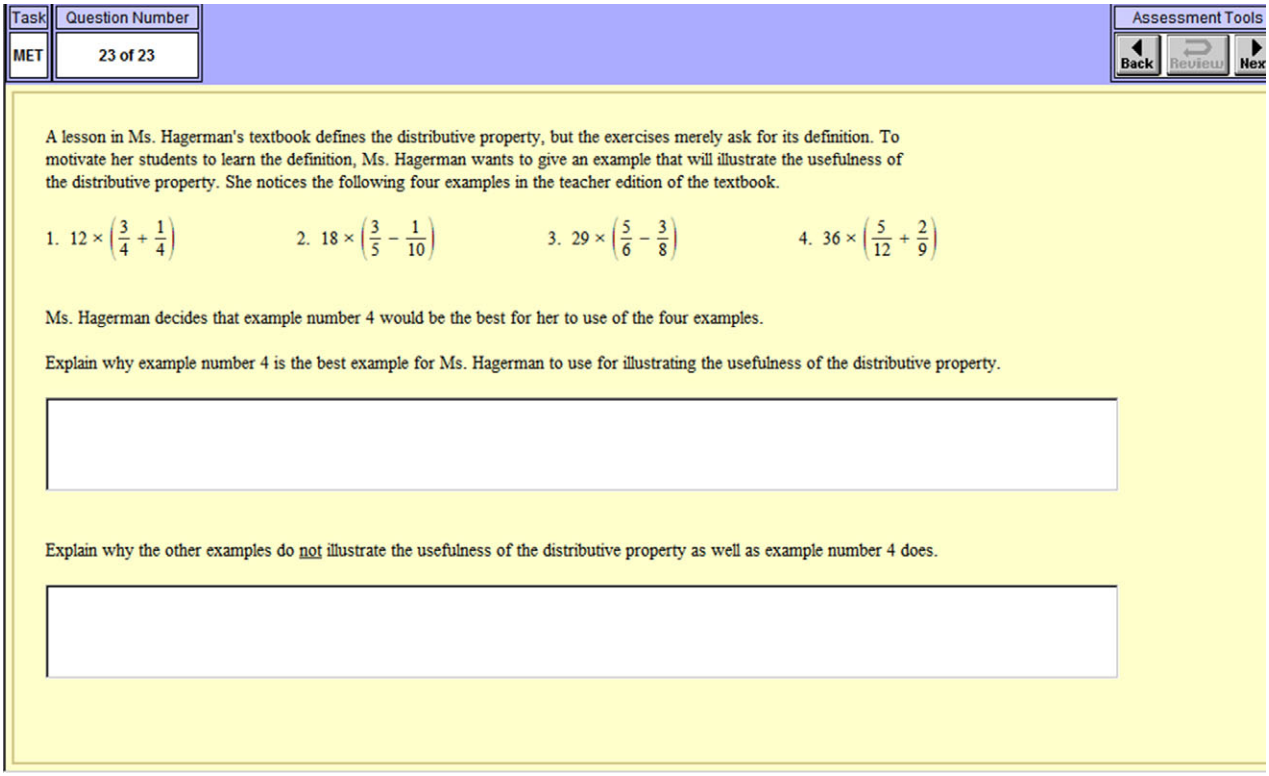



**Figure C23** Question #23.

## Appendix D

## English Language Arts (ELA) 7–9 Assessment

**Table D1** Item/Question Sequence Number for English Language Arts (ELA) 7–9 Assessment

| Item/question sequence number | Answer key | Item/question sequence number | Answer key |
|---|---|---|---|
| 1 | A | 15 | Lea, Akeem, Sarah |
| 2 | A | 16a | Literal |
| 3a | Not likely | 16b | Interpretive |
| 3b | Not likely | 16c | Personal |
| 3c | Likely | 17 | A |
| 3d | Not likely | 18 | Open-ended response |
| 4 | B | 19a | Will |
| 5 | A | 19b | Will |
| 6 | B | 19c | Will not |
| 7a | Should | 19d | Will not |
| 7b | Should not | 20 | B |
| 7c | Should | 21a | Has |
| 7d | Should not | 21b | Has not |
| 8 | D | 21c | Has |
| 9 | C | 22 | D |
| 10 | D | 23a | Does not describe |
| 11a | Accurate | 23b | Does not describe |
| 11b | Accurate | 23c | Describes |
| 11c | Not accurate | 23d | Describes |
| 11d | Accurate | 24a | Likely |
| 11e | Not accurate | 24b | Not likely |
| 11f | Not accurate | 24c | Likely |
| 12 | Open-ended response | 24d | Likely |
| 13 | D | 25 | A |
| 14 | B | 26 | B |

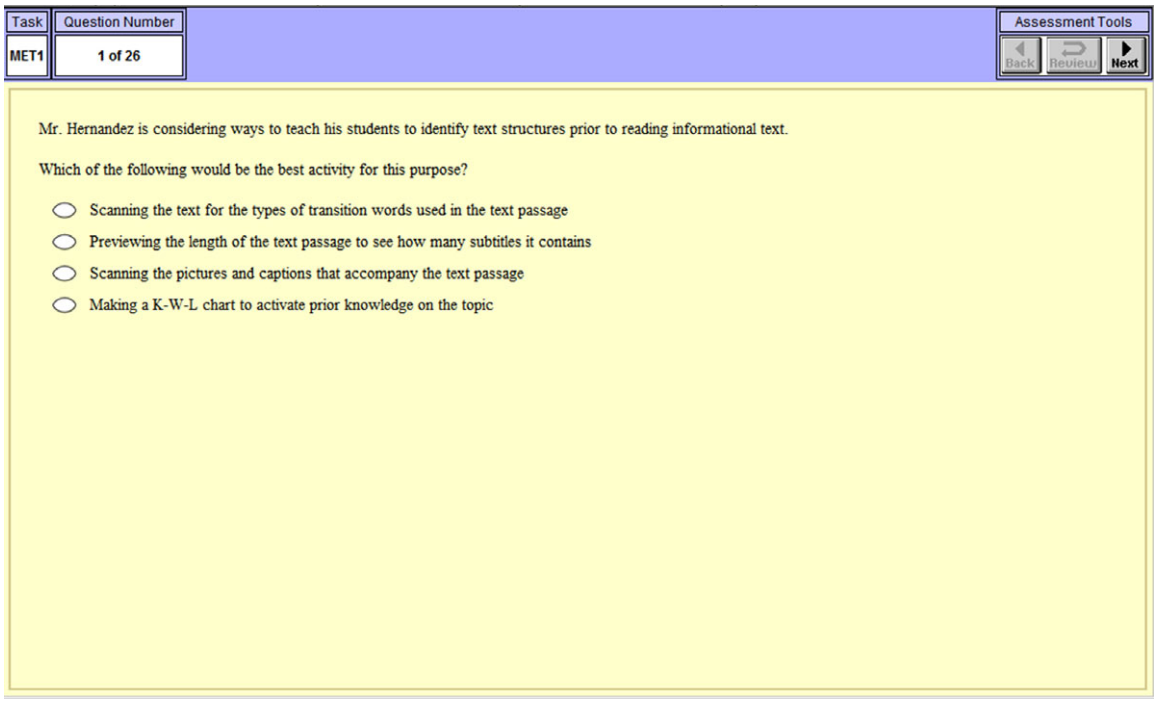



**Figure D1** Question #1.

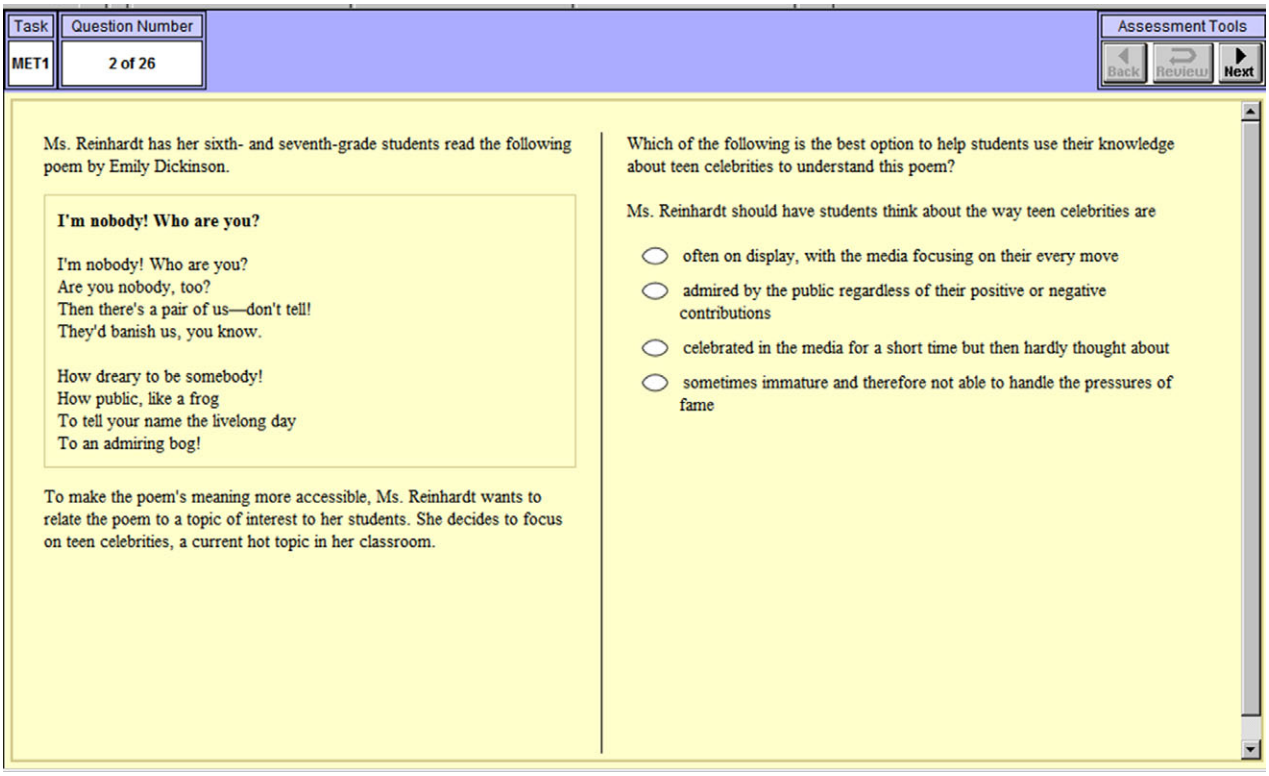



**Figure D2** Question #2. *I'm Nobody* by Emily Dickinson. Reprinted by permission of Dover Publications.

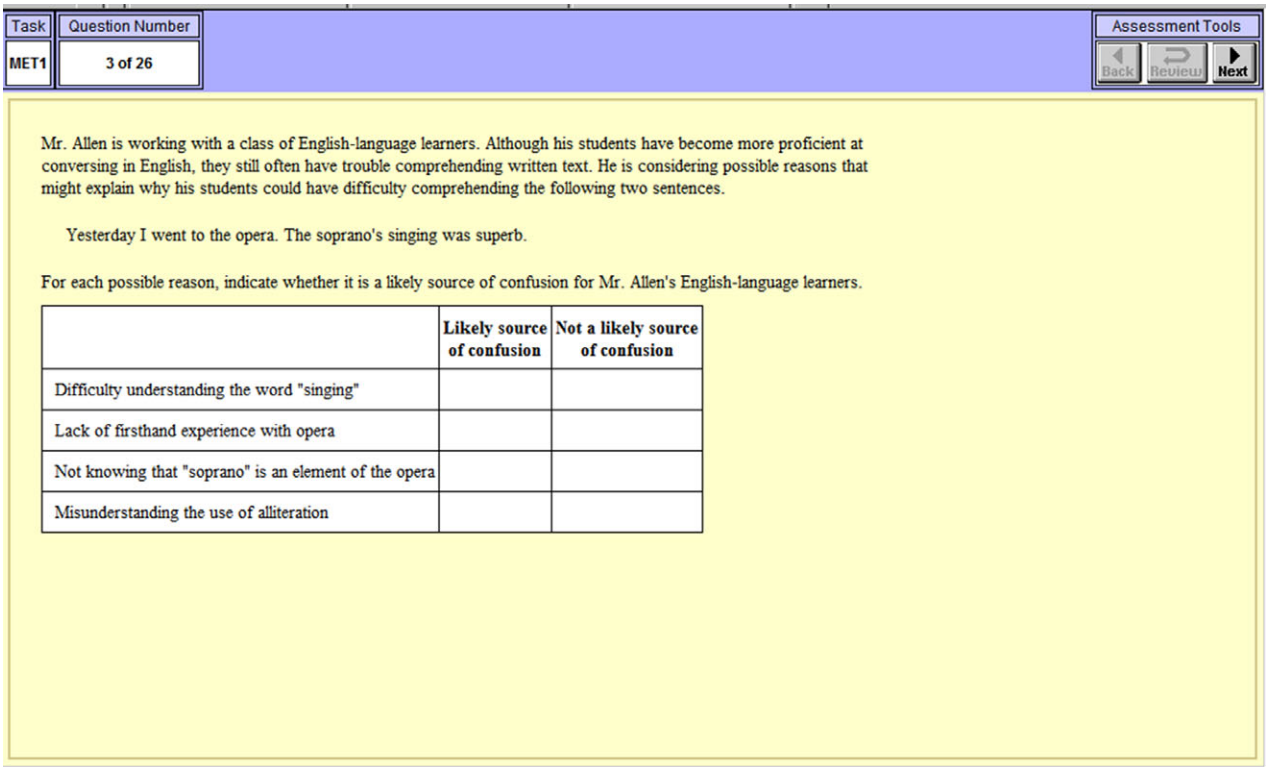



**Figure D3** Question #3.

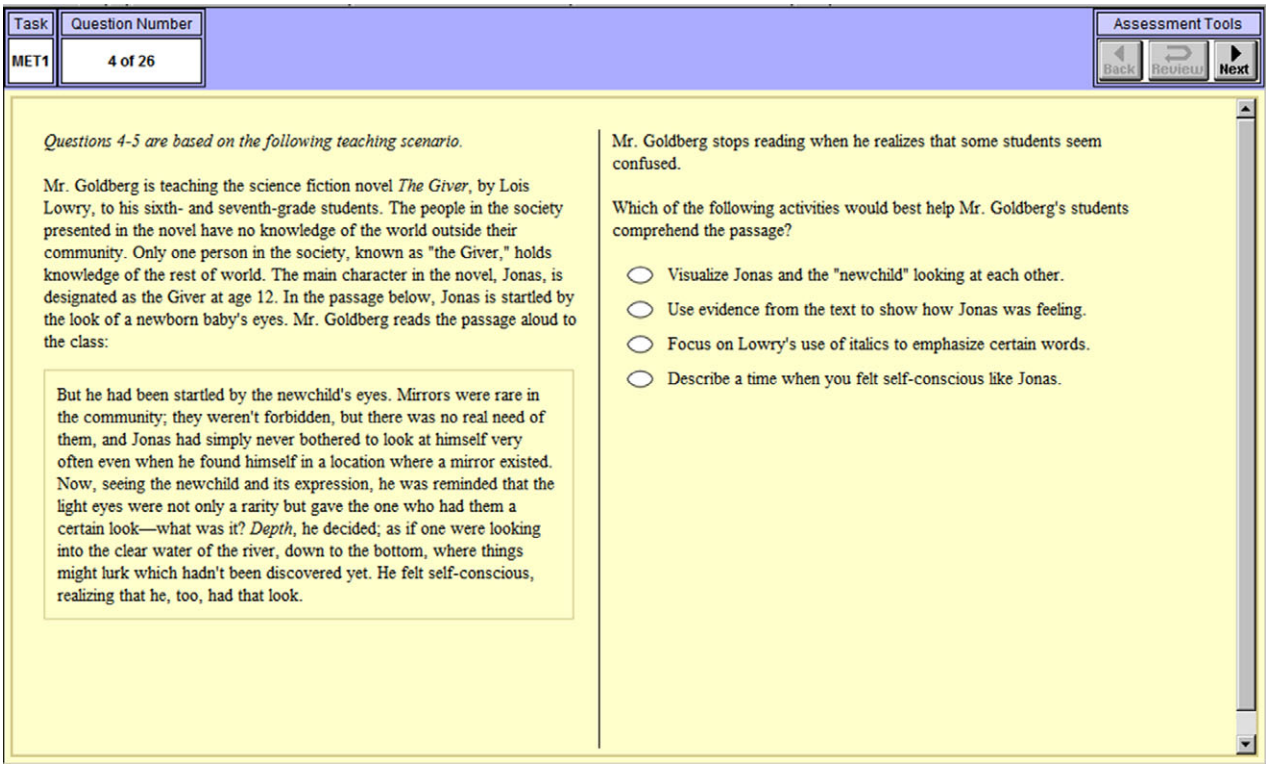



**Figure D4** Question #4.

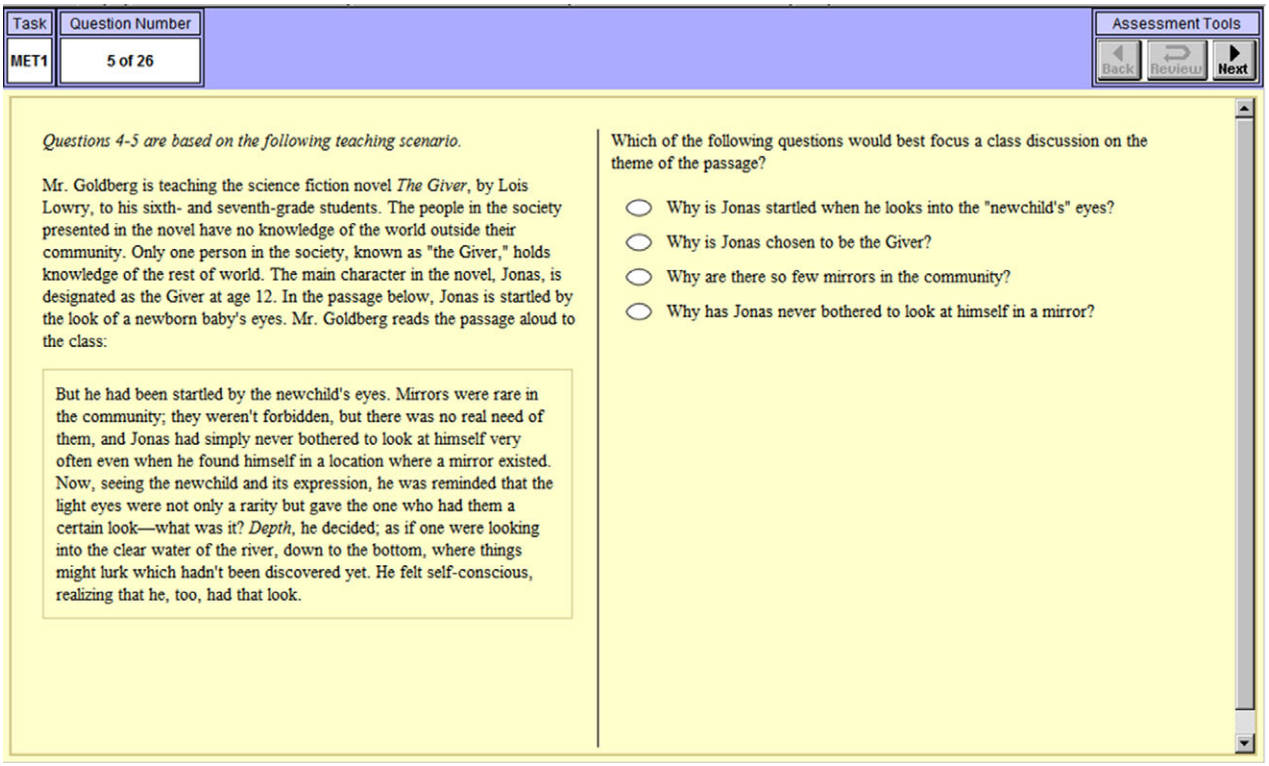



**Figure D5** Question #5.

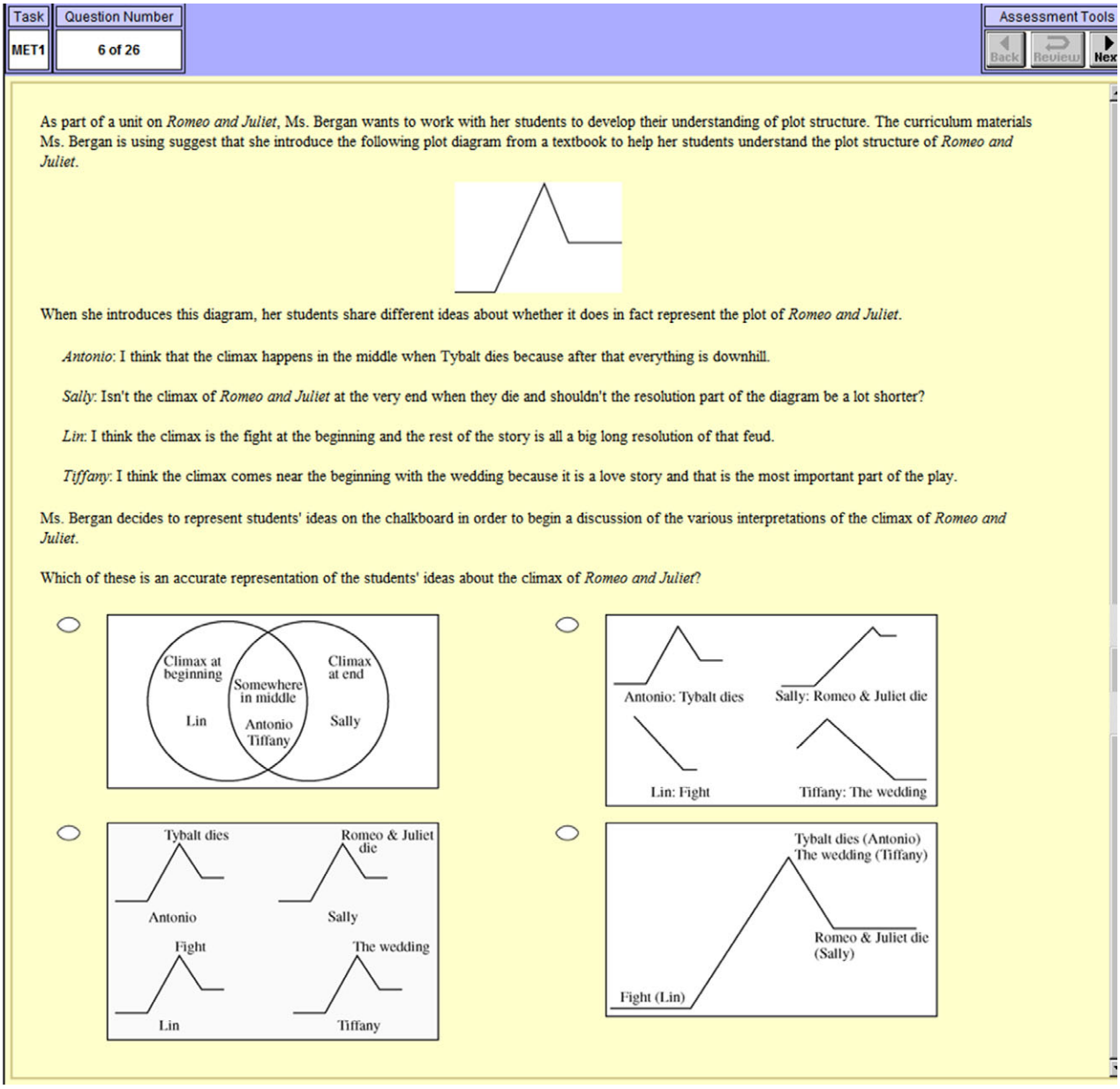



**Figure D6** Question #6.

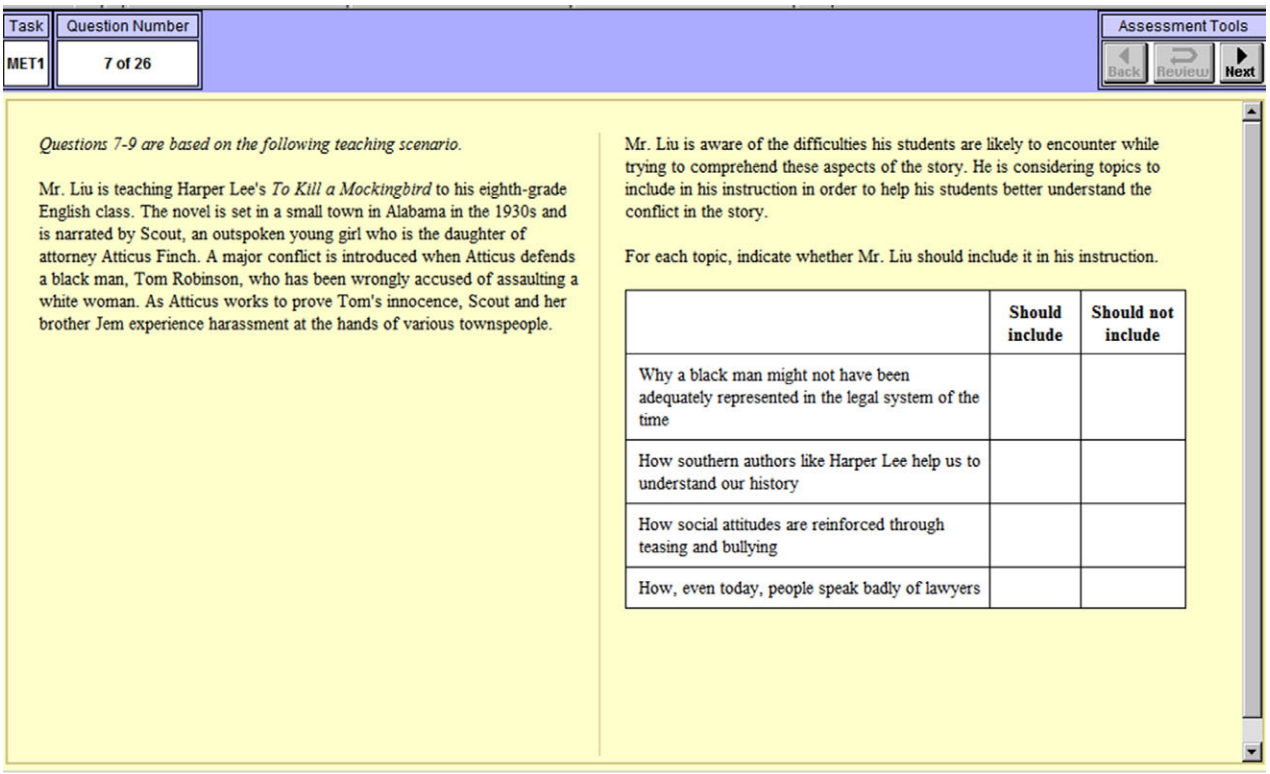



**Figure D7** Question #7.

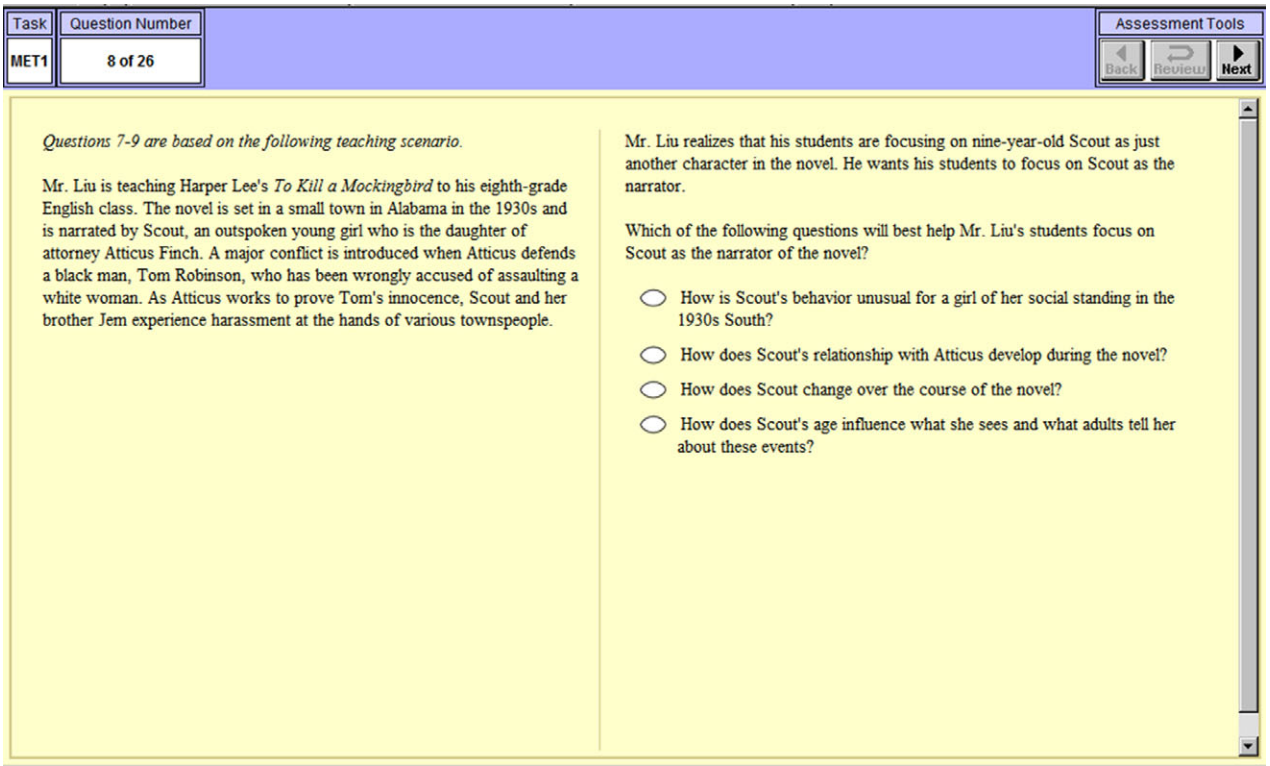



**Figure D8** Question #8.

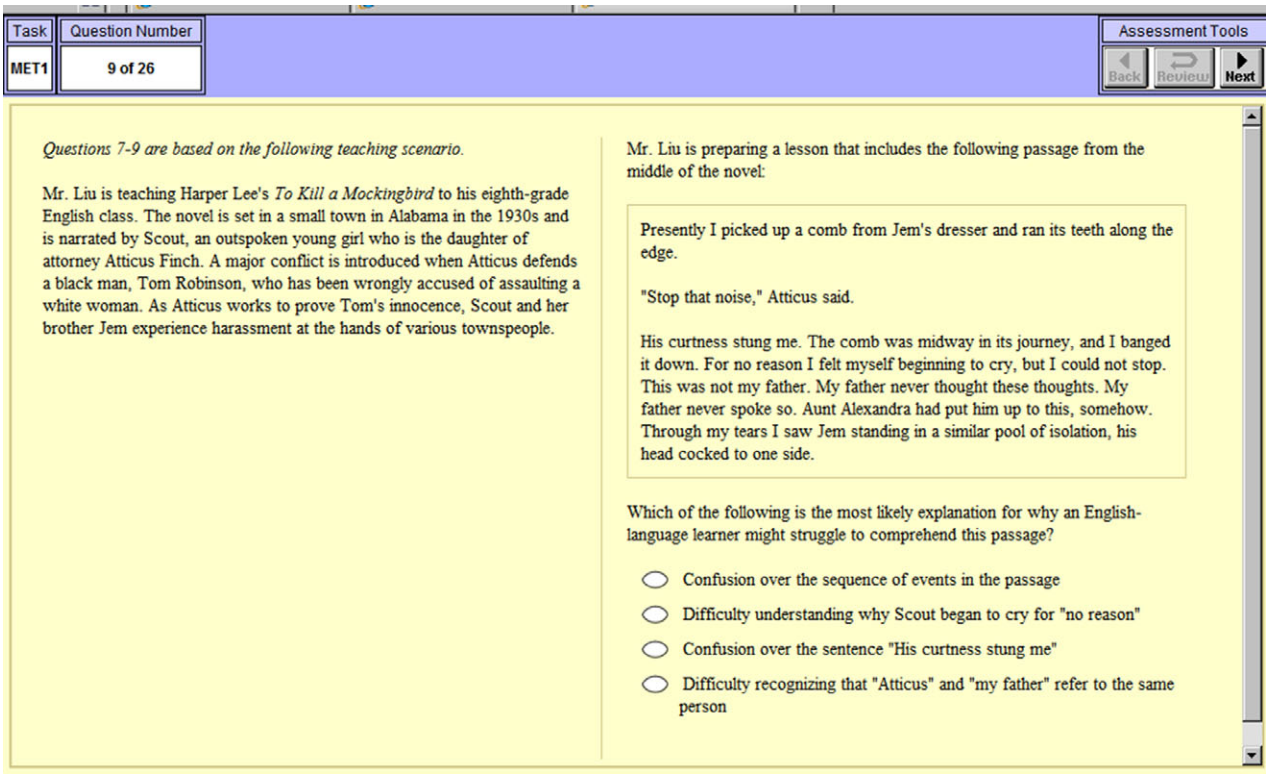



**Figure D9** Question #9.

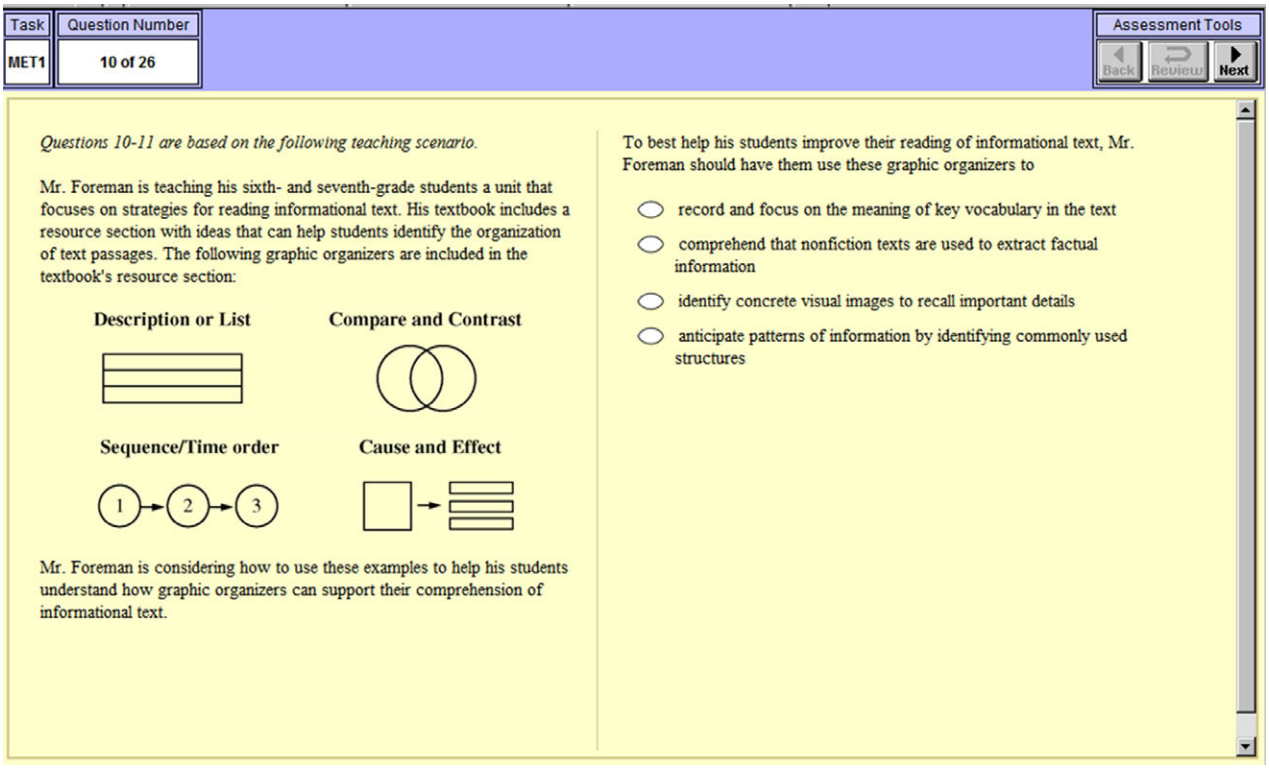



**Figure D10** Question #10.

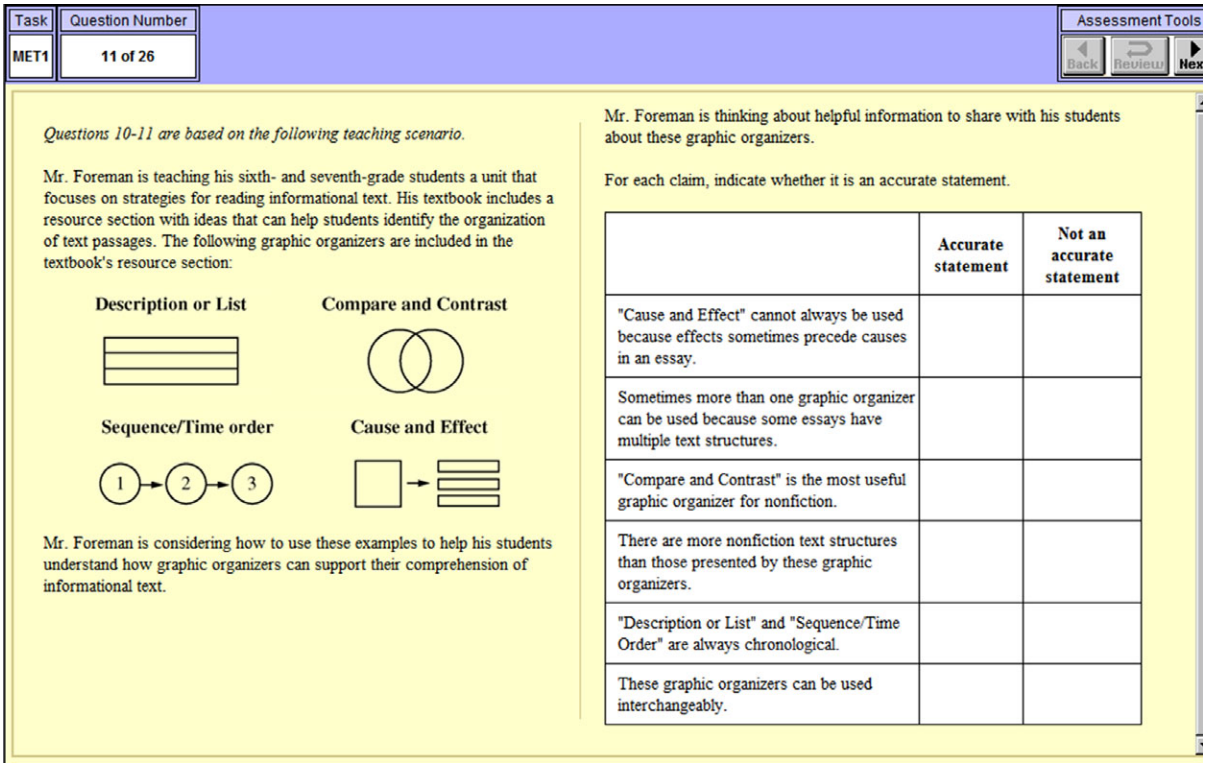



**Figure D11** Question #11.

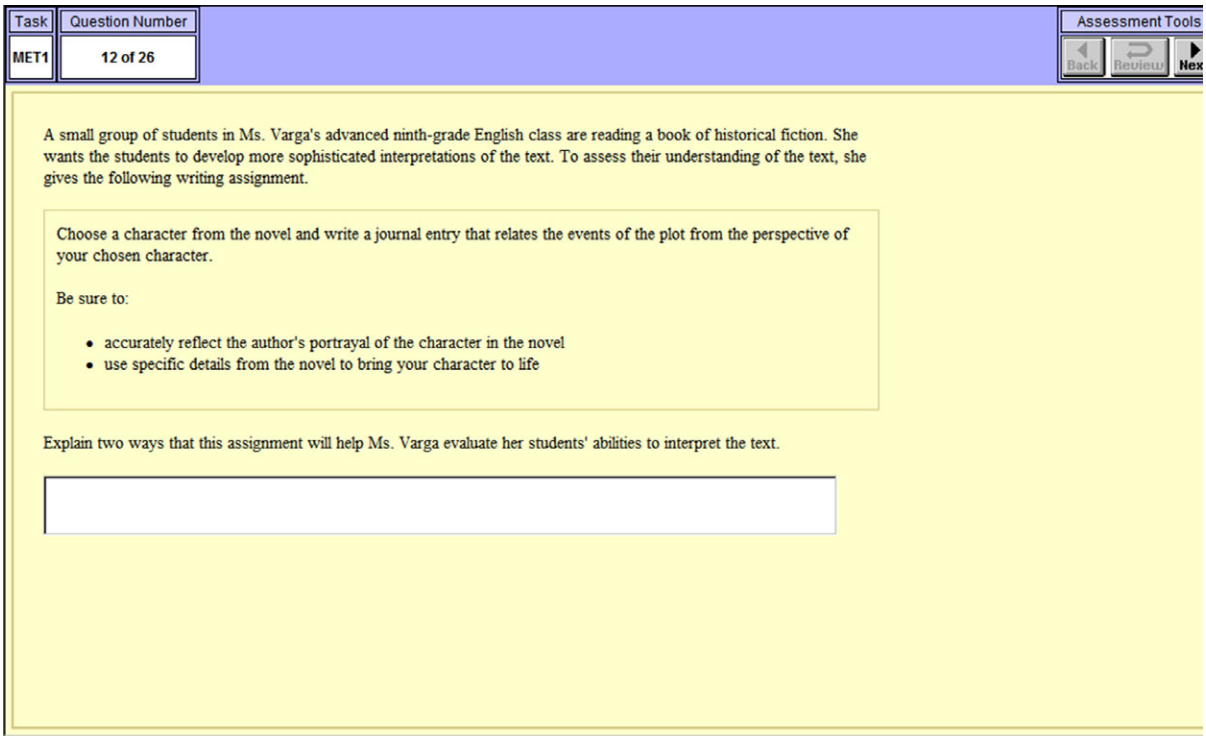



**Figure D12** Question #12.

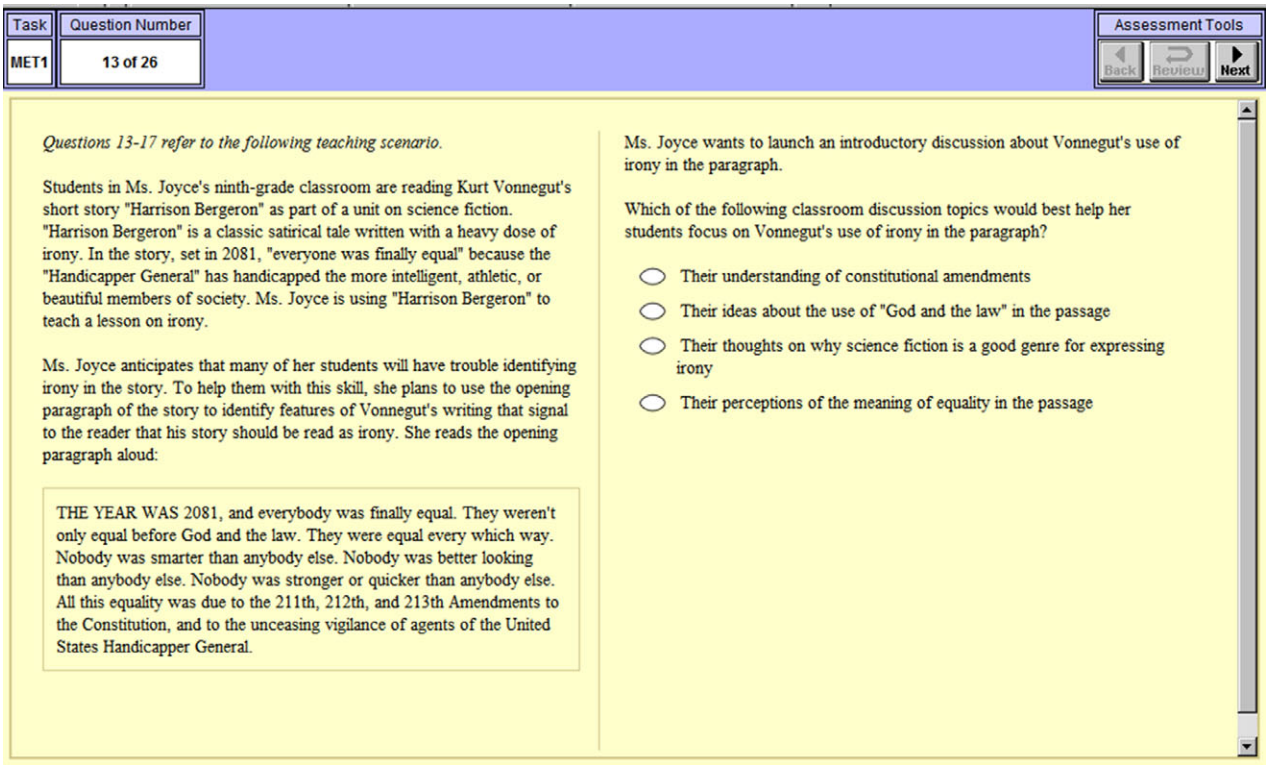



**Figure D13** Question #13. Adapted from *Welcome to the Monkey House* by Kurt Vonnegut, Jr. Copyright © 1961 by Kurt Vonnegut, Jr. Used by permission of Dell Publishing, a division of Random House, Inc.

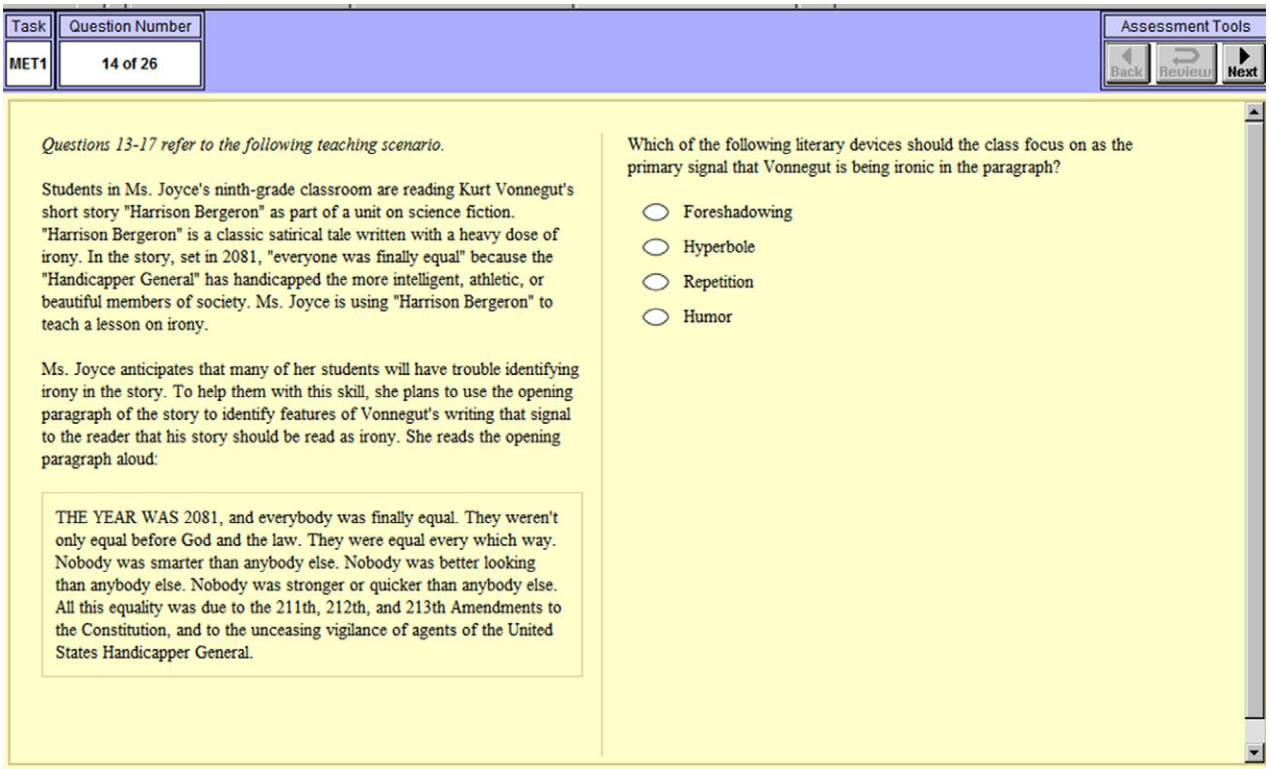



**Figure D14** Question #14. Adapted from *Welcome to the Monkey House* by Kurt Vonnegut, Jr. Copyright © 1961 by Kurt Vonnegut, Jr. Used by permission of Dell Publishing, a division of Random House, Inc.

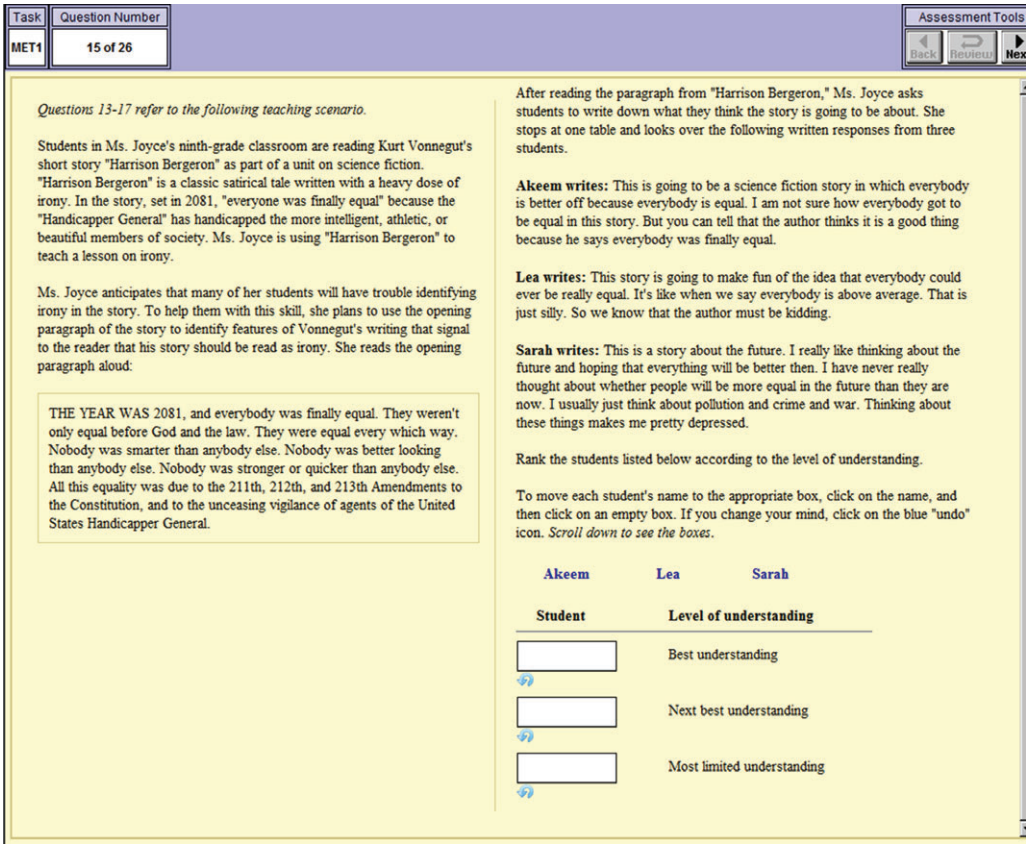



**Figure D15** Question #15. Adapted from *Welcome to the Monkey House* by Kurt Vonnegut, Jr. Copyright © 1961 by Kurt Vonnegut, Jr. Used by permission of Dell Publishing, a division of Random House, Inc.

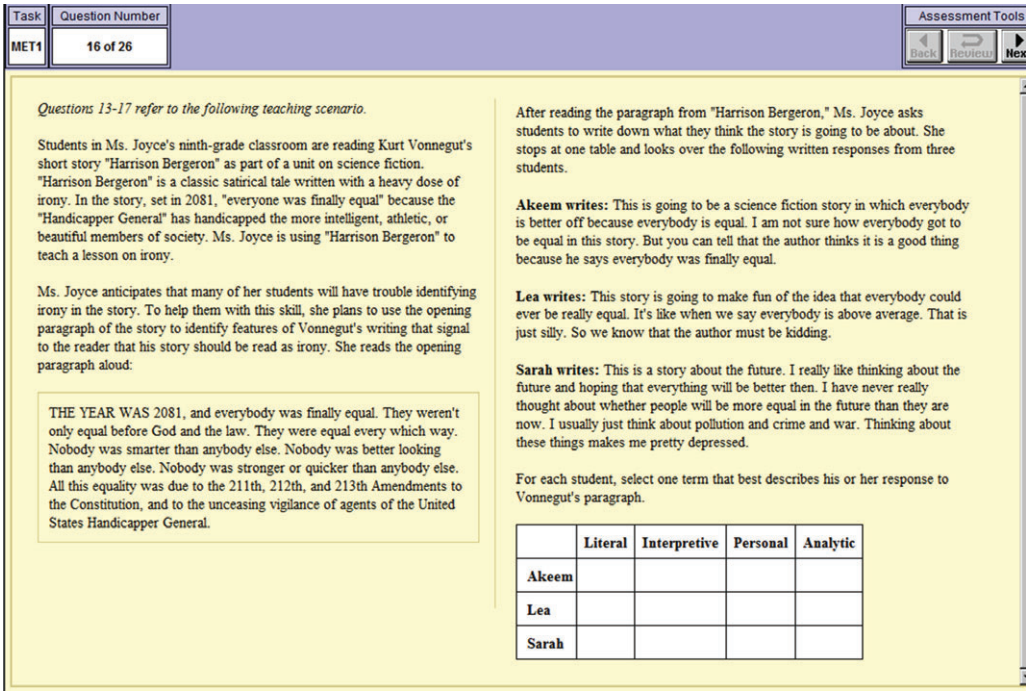



**Figure D16** Question #16. Adapted from *Welcome to the Monkey House* by Kurt Vonnegut, Jr. Copyright © 1961 by Kurt Vonnegut, Jr. Used by permission of Dell Publishing, a division of Random House, Inc.

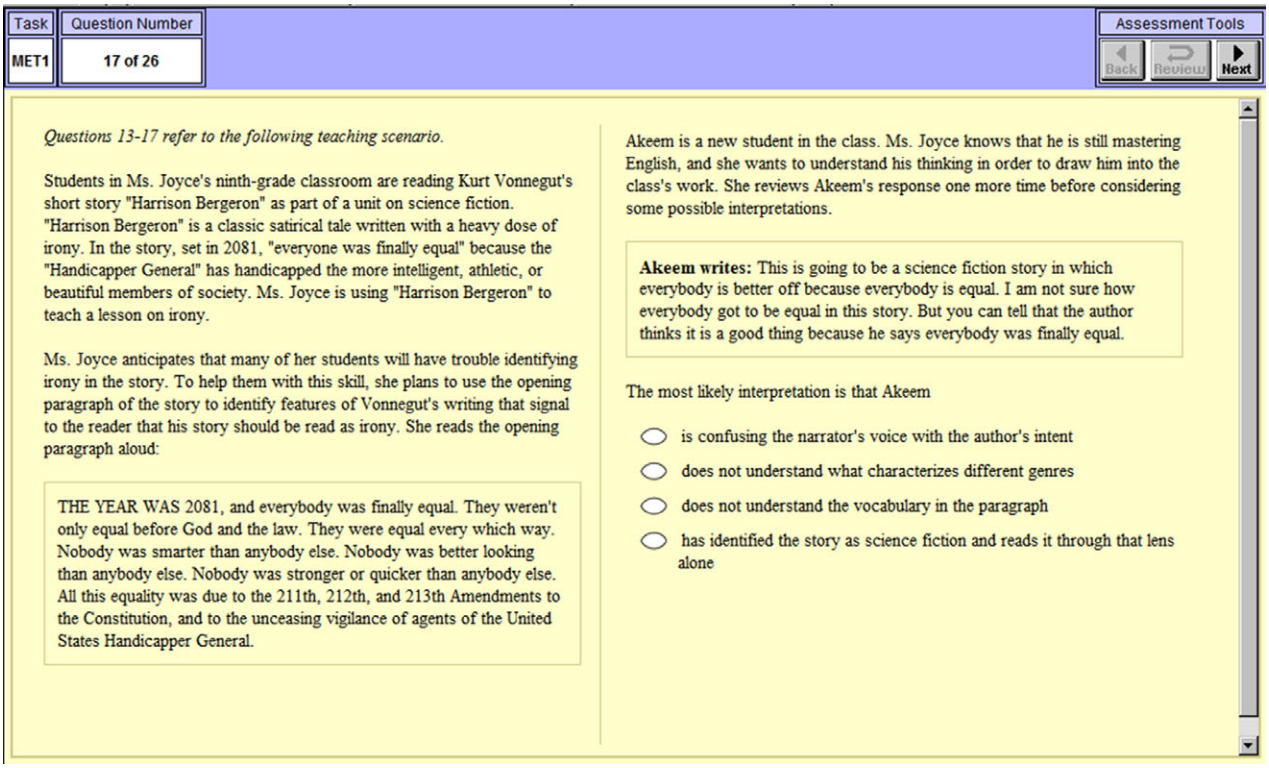



**Figure D17** Question #17. Adapted from *Welcome to the Monkey House* by Kurt Vonnegut, Jr. Copyright © 1961 by Kurt Vonnegut, Jr. Used by permission of Dell Publishing, a division of Random House, Inc.

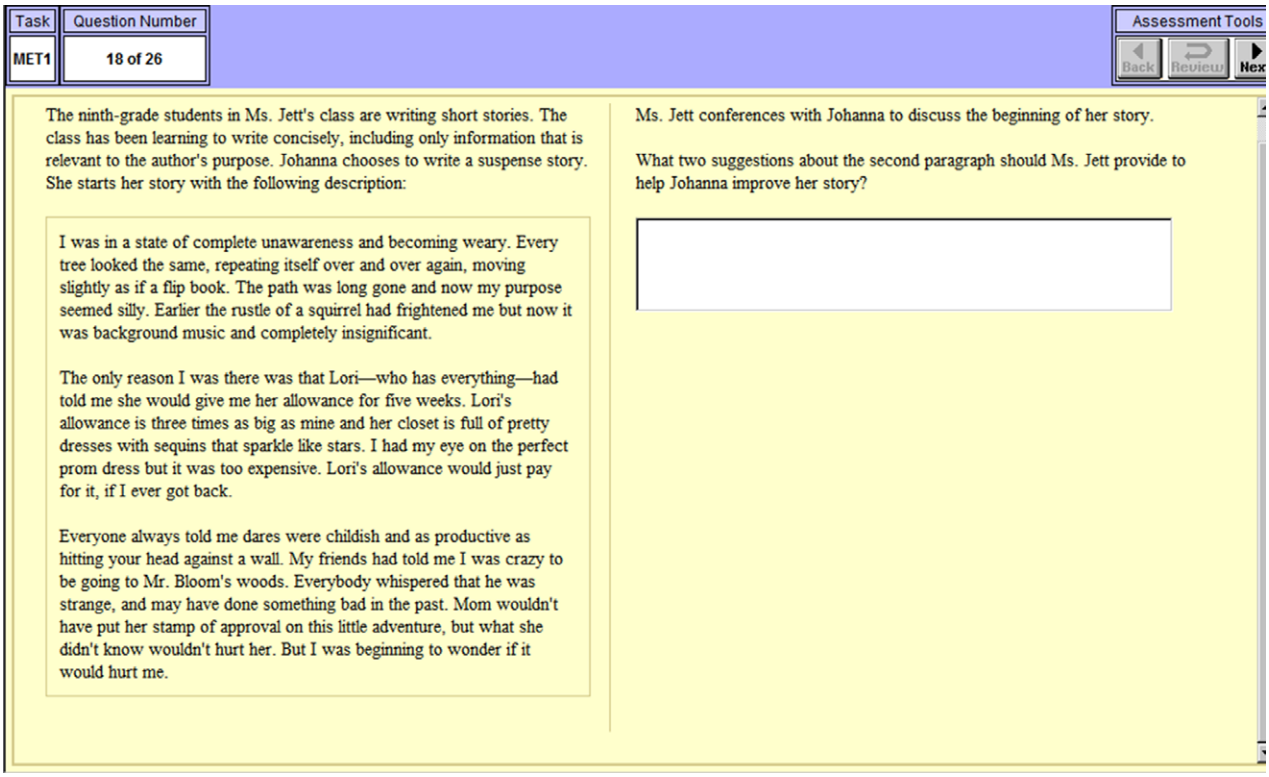
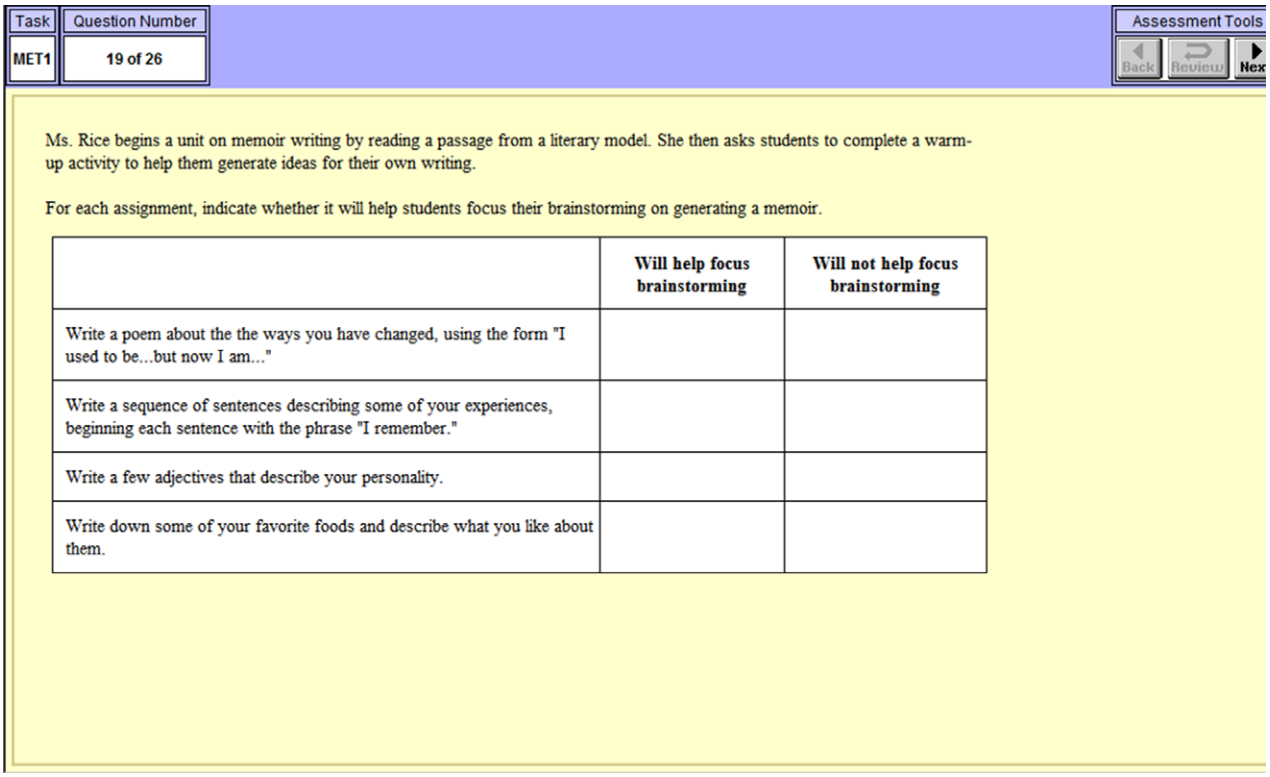



**Figure D18** Question #18.

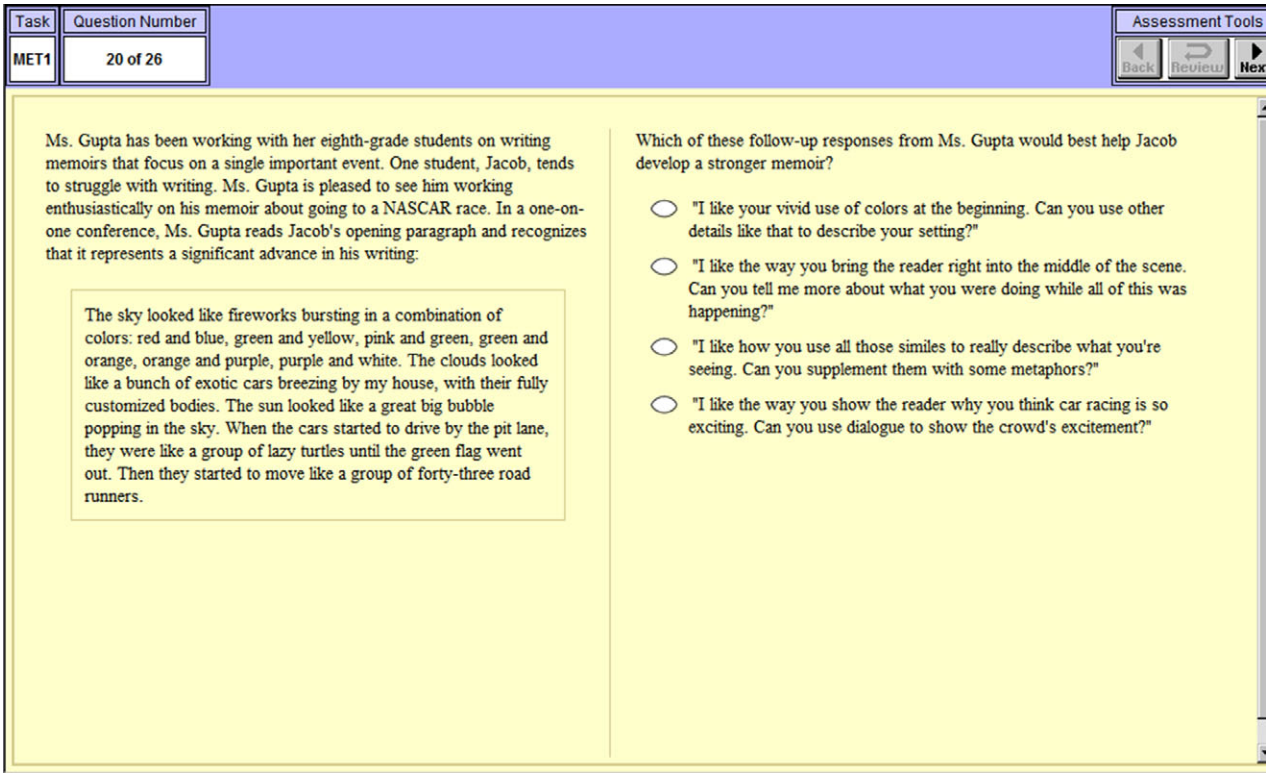



Ms. Rice begins a unit on memoir writing by reading a passage from a literary model. She then asks students to complete a warm-up activity to help them generate ideas for their own writing.

For each assignment, indicate whether it will help students focus their brainstorming on generating a memoir.

| | **Will help focus brainstorming** | **Will not help focus brainstorming** |
|---|---|---|
| Write a poem about the the ways you have changed, using the form "I used to be...but now I am..." | | |
| Write a sequence of sentences describing some of your experiences, beginning each sentence with the phrase "I remember." | | |
| Write a few adjectives that describe your personality. | | |
| Write down some of your favorite foods and describe what you like about them. | | |

**Figure D19** Question #19.




Ms. Gupta has been working with her eighth-grade students on writing memoirs that focus on a single important event. One student, Jacob, tends to struggle with writing. Ms. Gupta is pleased to see him working enthusiastically on his memoir about going to a NASCAR race. In a one-on-one conference, Ms. Gupta reads Jacob's opening paragraph and recognizes that it represents a significant advance in his writing:

> The sky looked like fireworks bursting in a combination of colors: red and blue, green and yellow, pink and green, green and orange, orange and purple, purple and white. The clouds looked like a bunch of exotic cars breezing by my house, with their fully customized bodies. The sun looked like a great big bubble popping in the sky. When the cars started to drive by the pit lane, they were like a group of lazy turtles until the green flag went out. Then they started to move like a group of forty-three road runners.

Which of these follow-up responses from Ms. Gupta would best help Jacob develop a stronger memoir?

- "I like your vivid use of colors at the beginning. Can you use other details like that to describe your setting?"
- "I like the way you bring the reader right into the middle of the scene. Can you tell me more about what you were doing while all of this was happening?"
- "I like how you use all those similes to really describe what you're seeing. Can you supplement them with some metaphors?"
- "I like the way you show the reader why you think car racing is so exciting. Can you use dialogue to show the crowd's excitement?"

**Figure D20** Question #20.

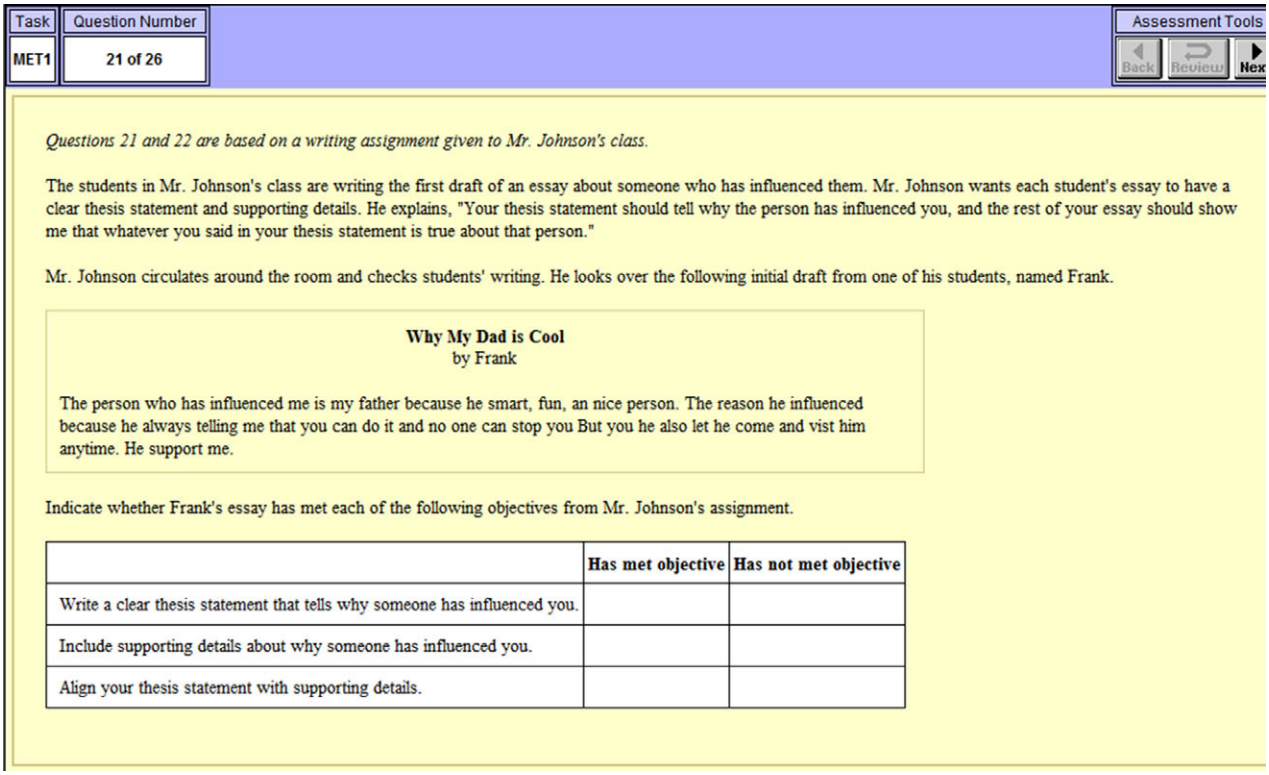



**Figure D21** Question #21.


**Figure D22** Question #22.

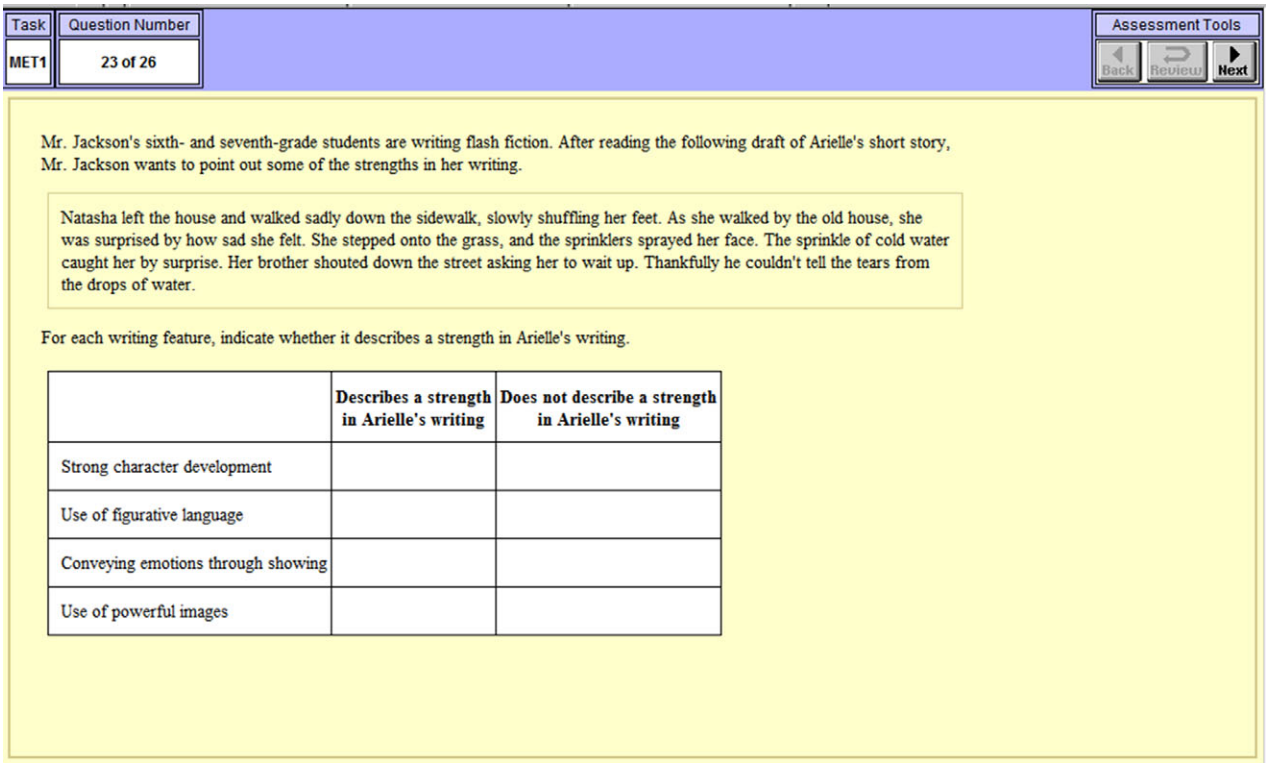



**Figure D23** Question #23.

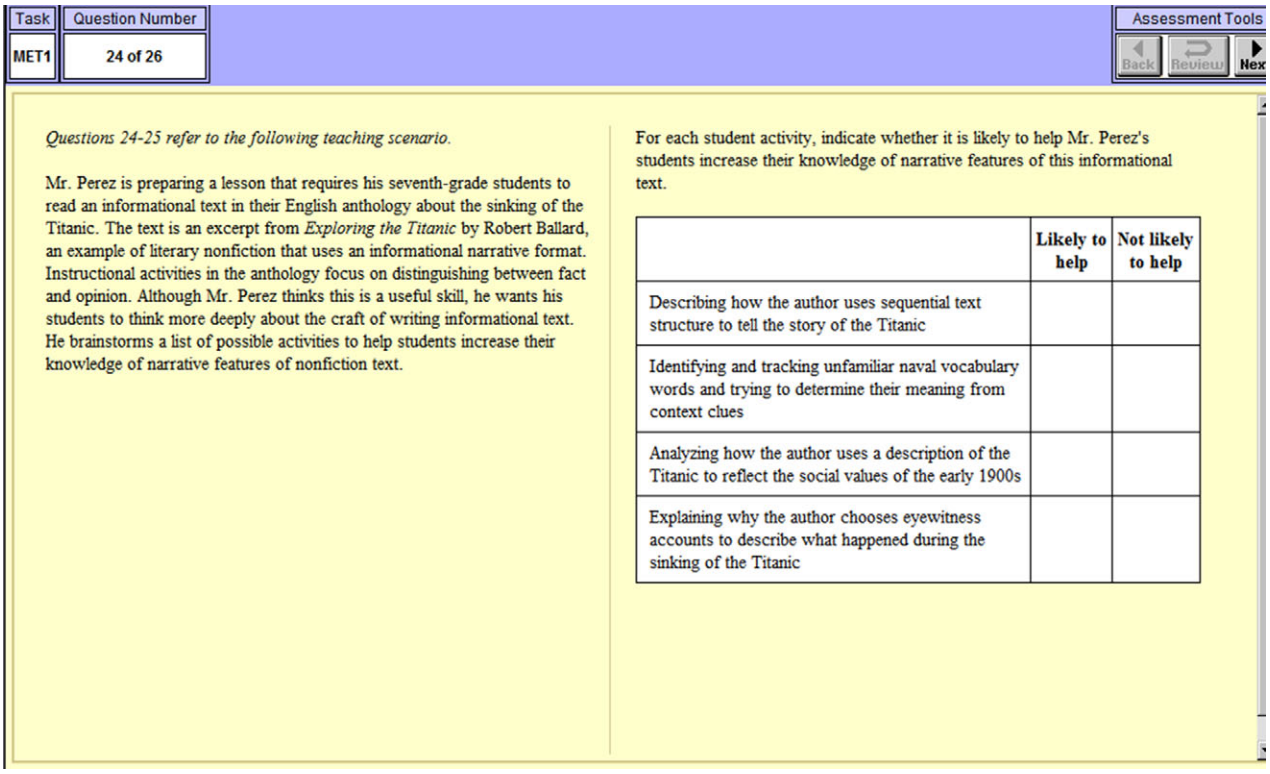



**Figure D24** Question #24.

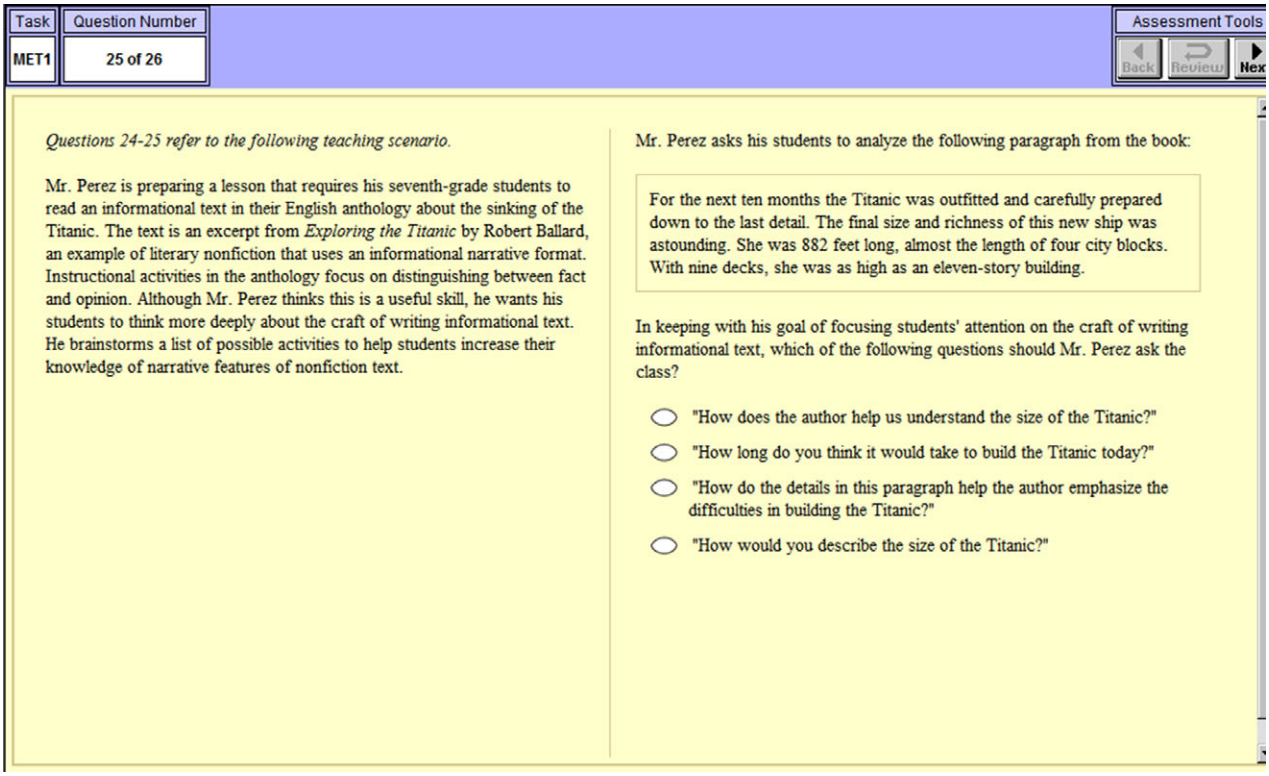



**Figure D25** Question #25.

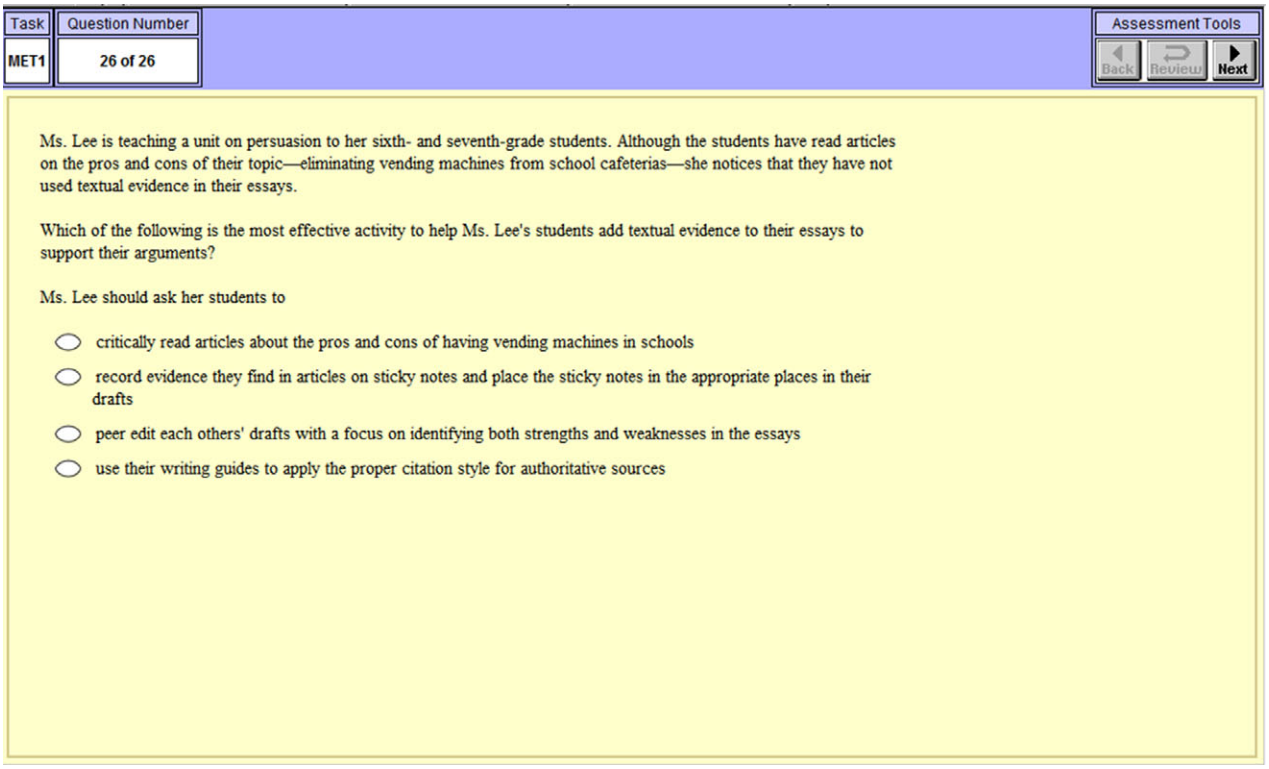



**Figure D26** Question #26.

# Appendix E

## Algebra I Assessment

**Table E1** Item/Question Sequence Number for Algebra I Assessment

| Item/question sequence number | Answer key | Item/question sequence number | Answer key |
|---|---|---|---|
| 1 | B | 11d | Valid |
| 2 | C | 11e | Not valid |
| 3a | Provides | 12 | B |
| 3b | Does not provide | 13 | B |
| 3c | Provides | 14 | C |
| 3d | Does not provide | 15a | Provides |
| 3e | Provides | 15b | Provides |
| 4 | C | 15c | Provides |
| 5 | C | 15d | Does not provide |
| 6a | Would not support | 15e | Does not provide |
| 6b | Would support | 16 | C |
| 6c | Would support | 17 | C |
| 7 | B | 18 | B |
| 8 | C | 19 | D |
| 9 | C | 20 | C |
| 10 | B | 21 | D |
| 11a | Valid | 22 | D |
| 11b | Valid | 23 | C |
| 11c | Not valid | | |

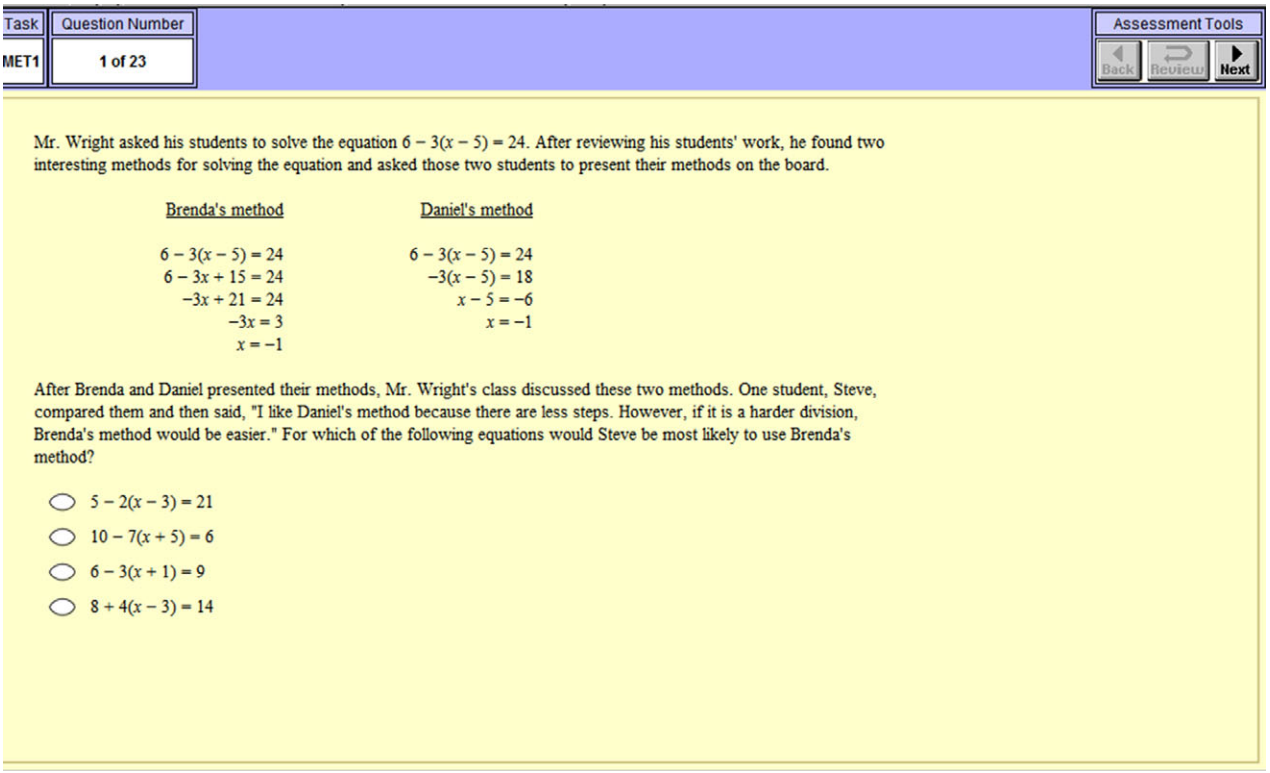



**Figure E1** Question #1.

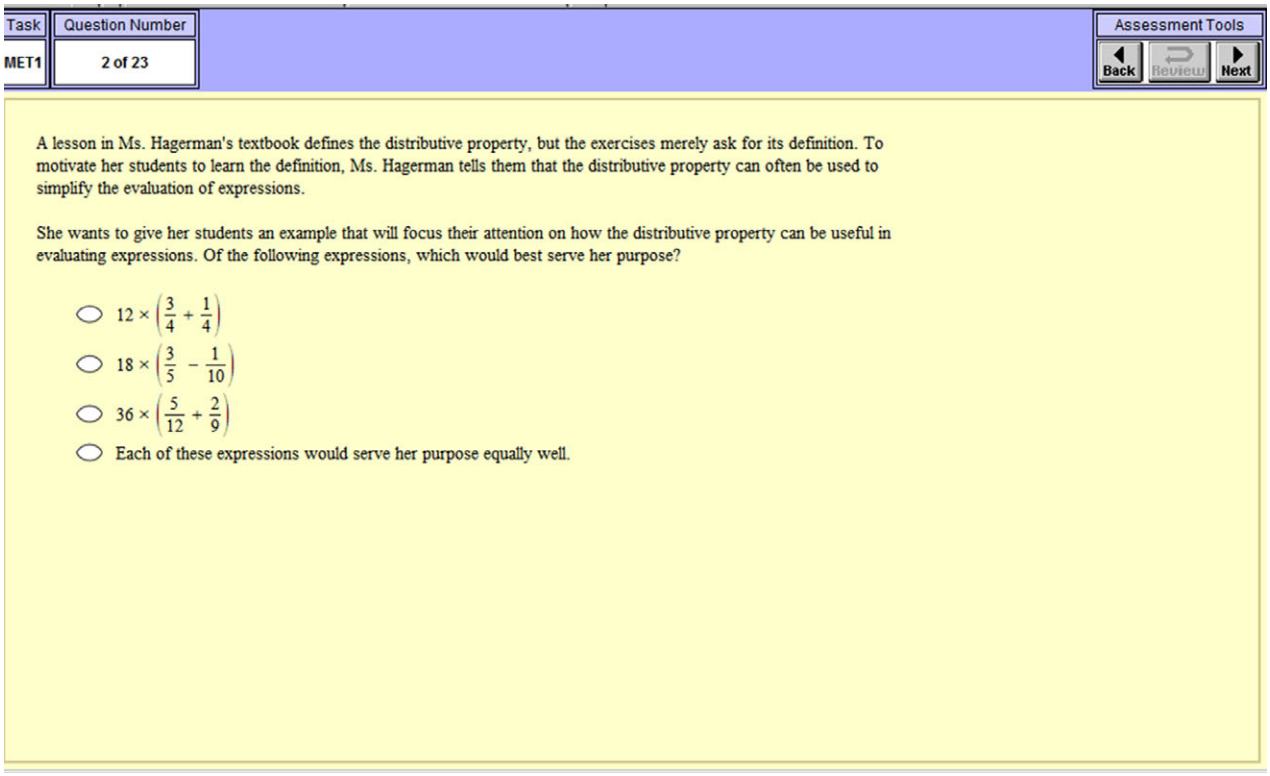



**Figure E2** Question #2.

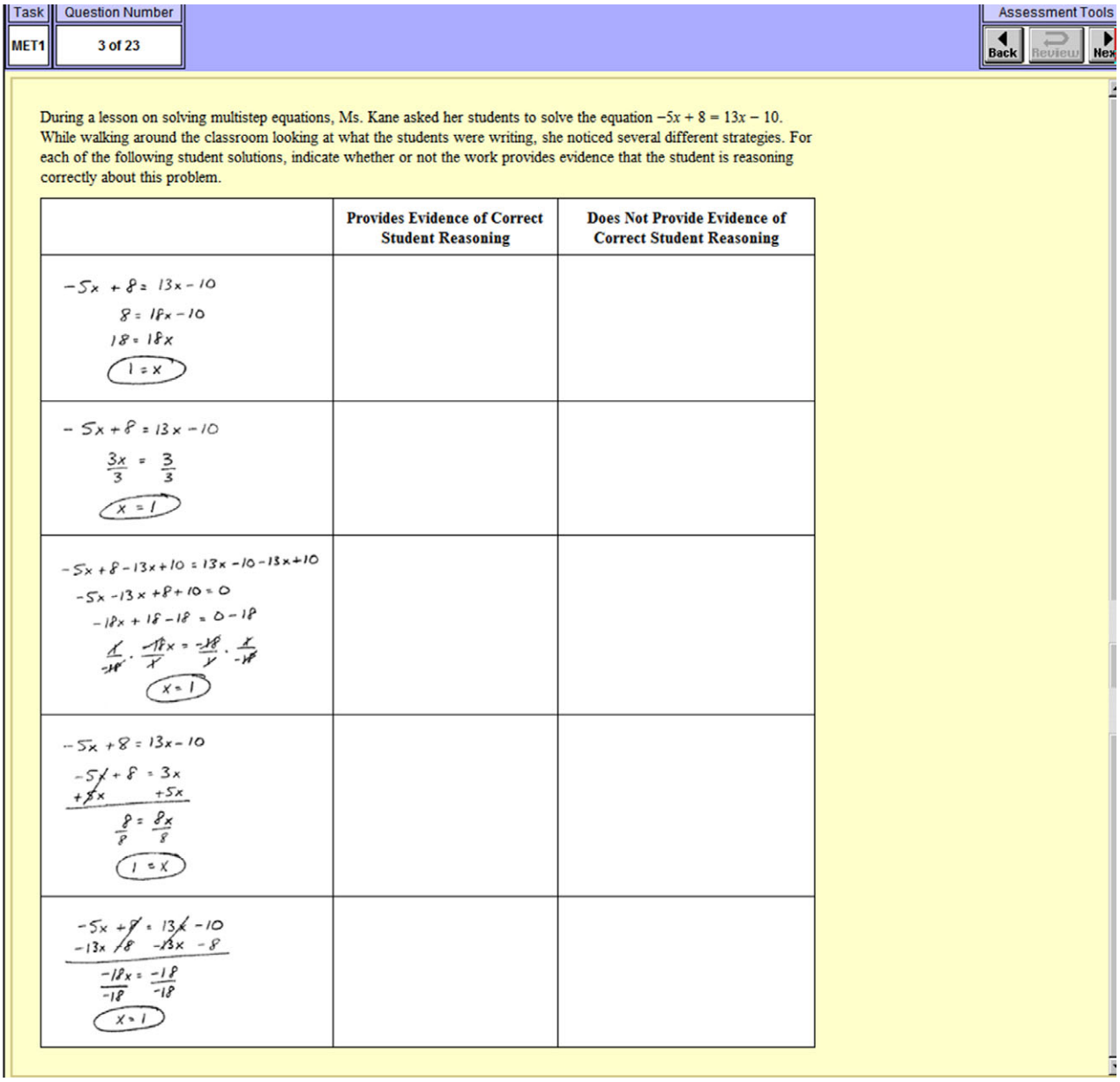



**Figure E3** Question #3.

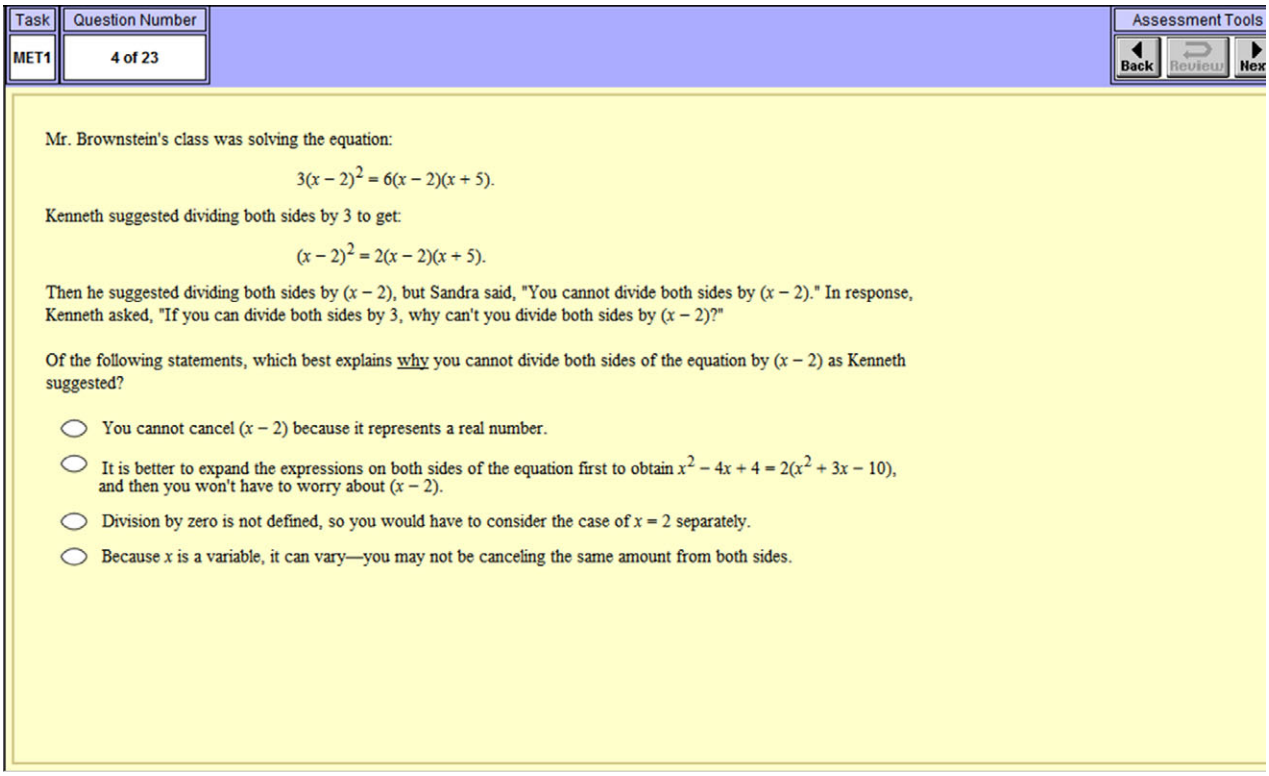



**Figure E4** Question #4.

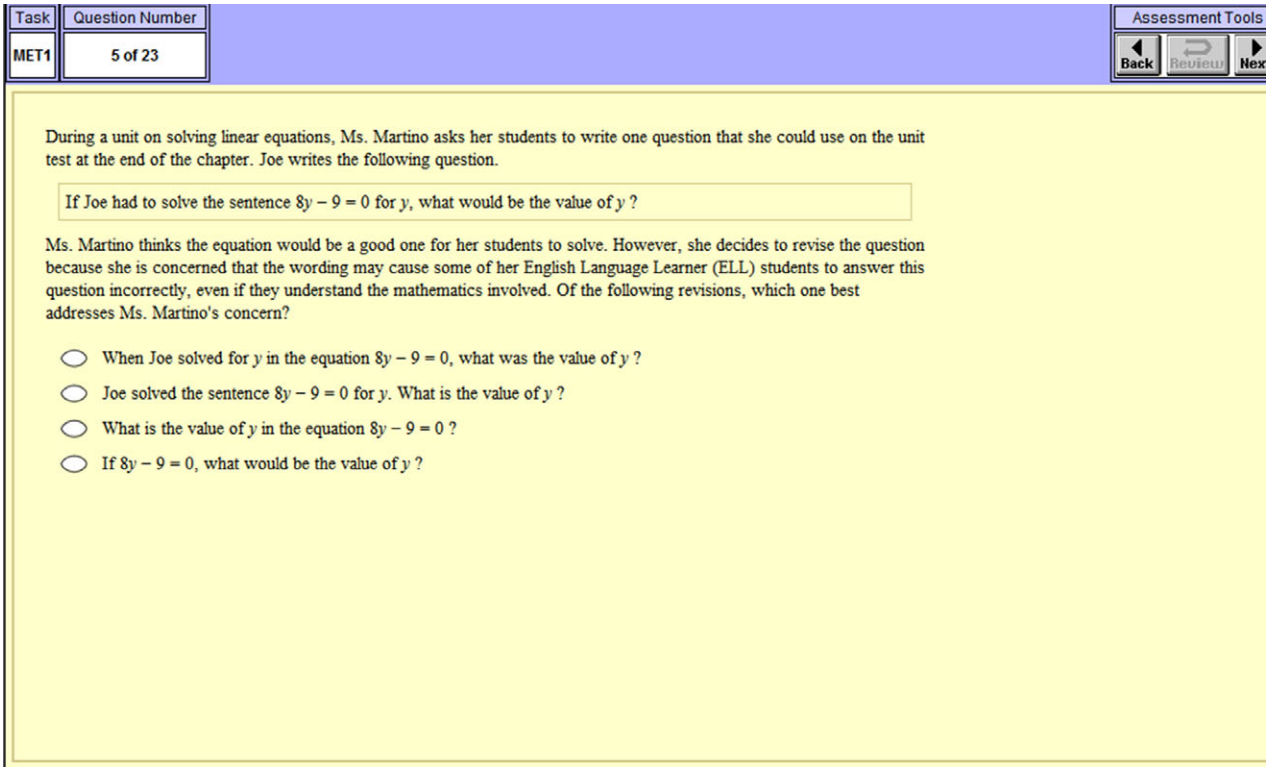



**Figure E5** Question #5.

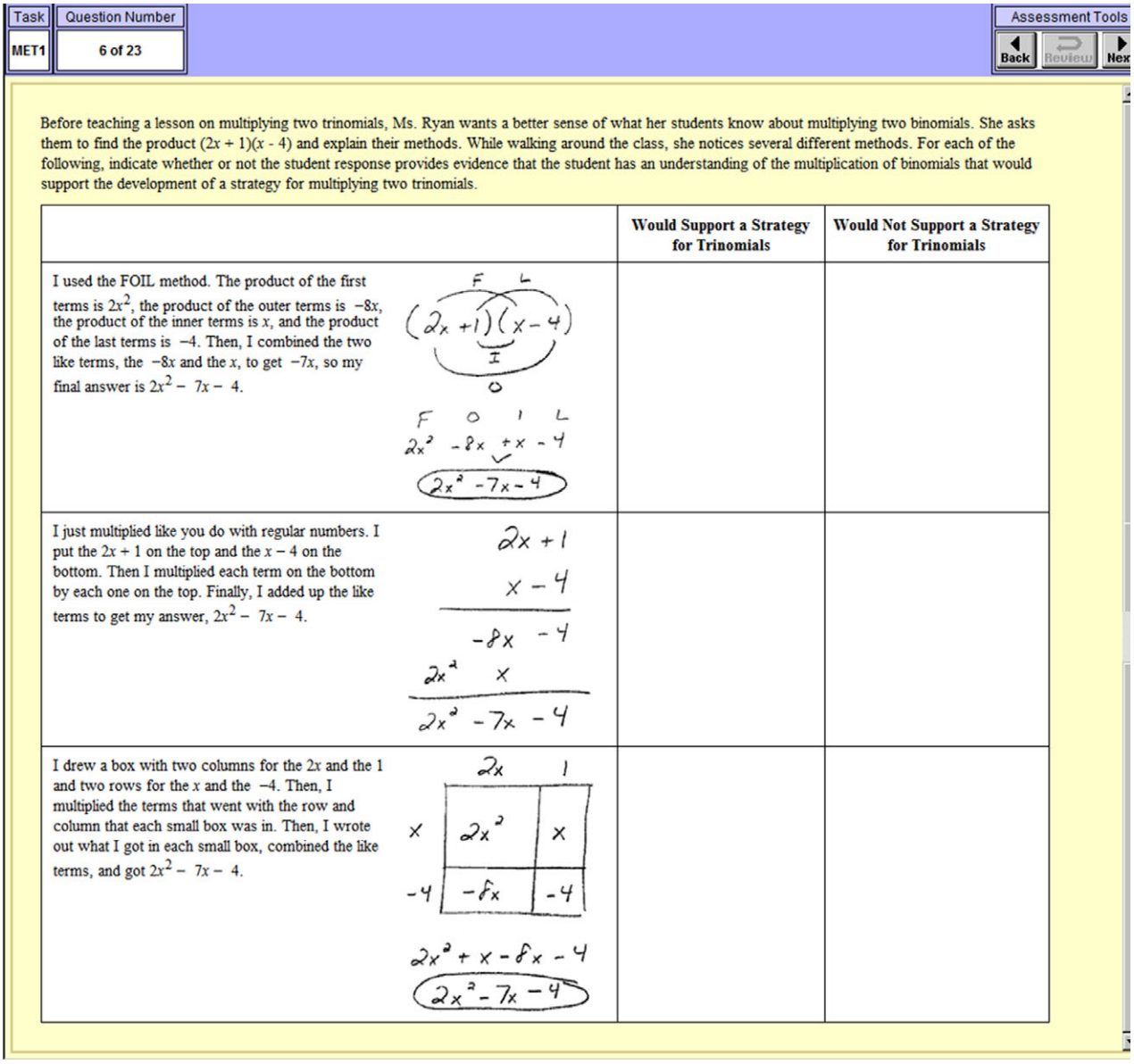



**Figure E6** Question #6.

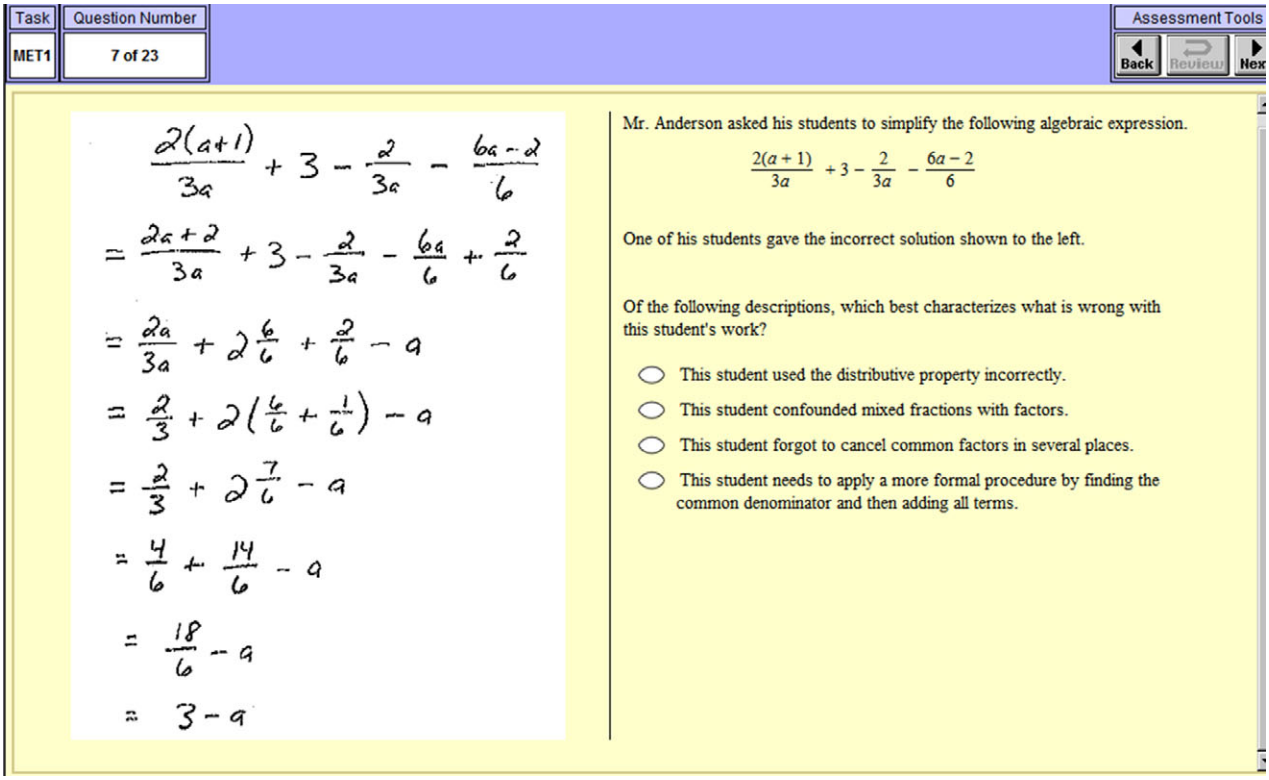



**Figure E7** Question #7.

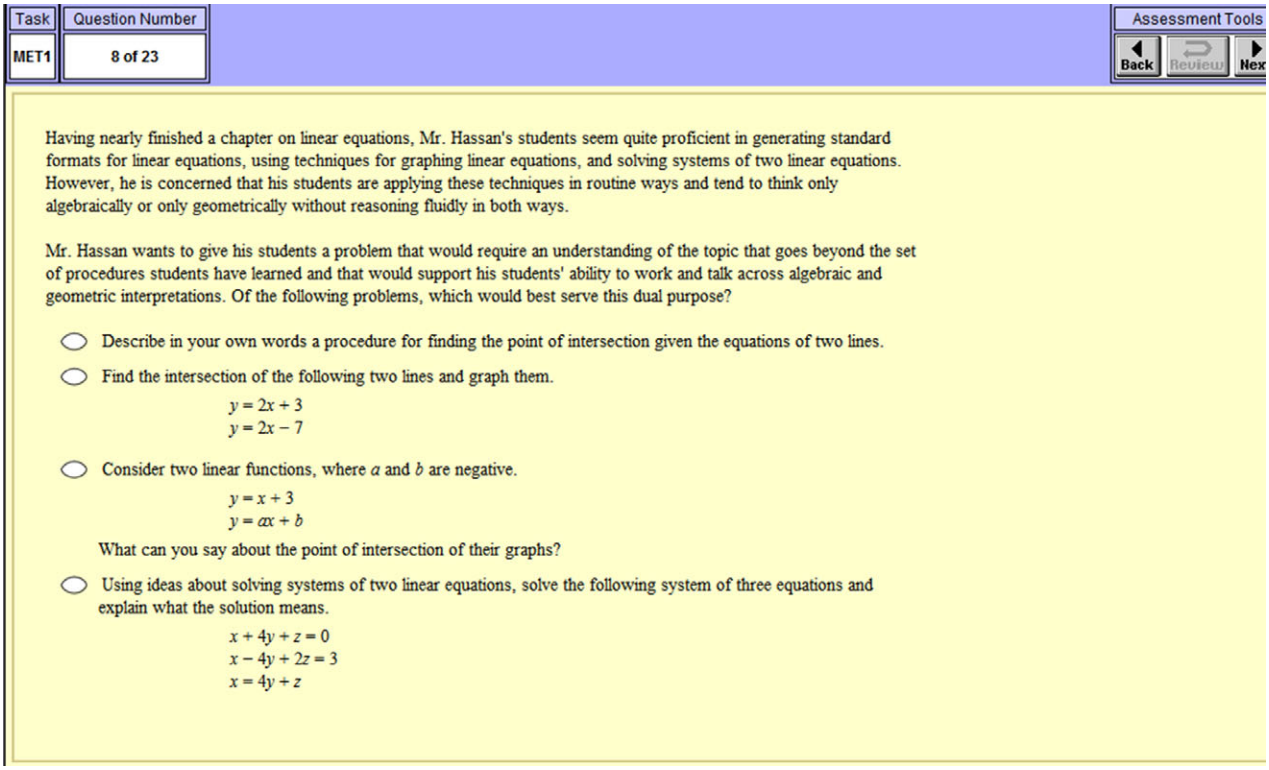



**Figure E8** Question #8.

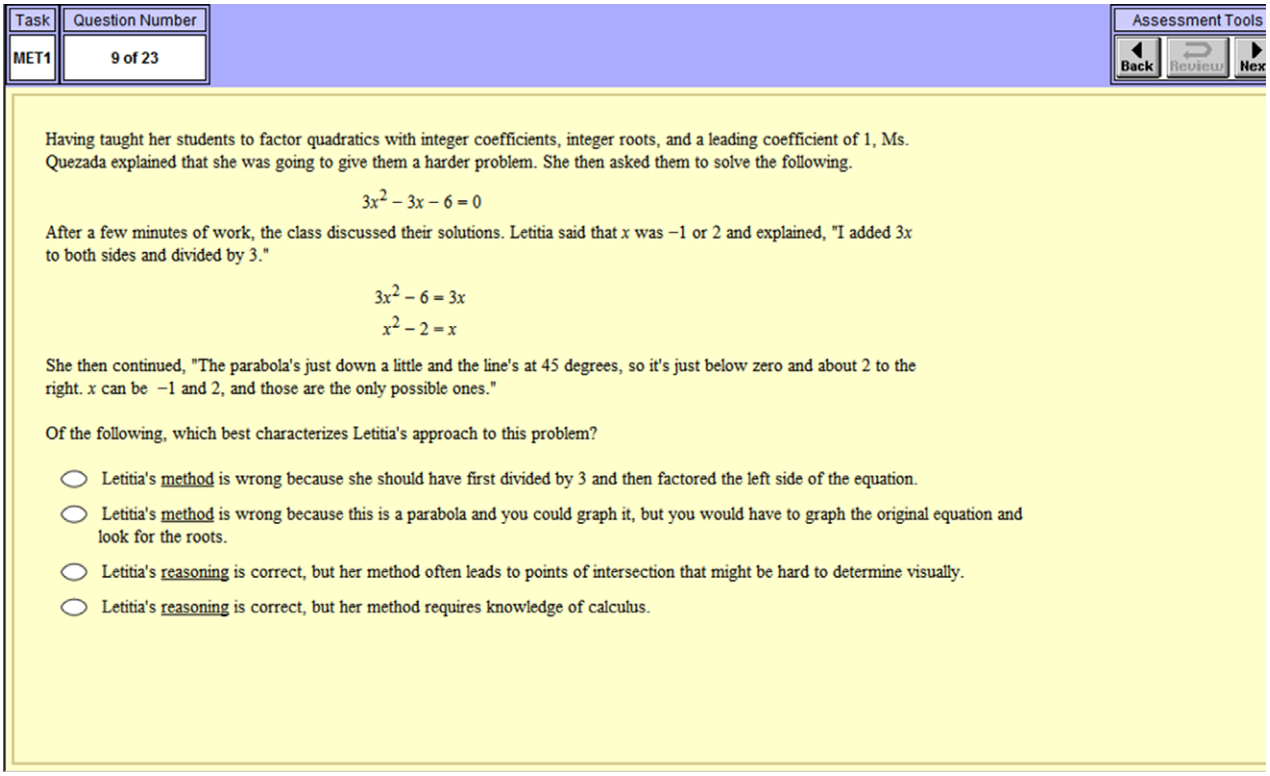



**Figure E9** Question #9.

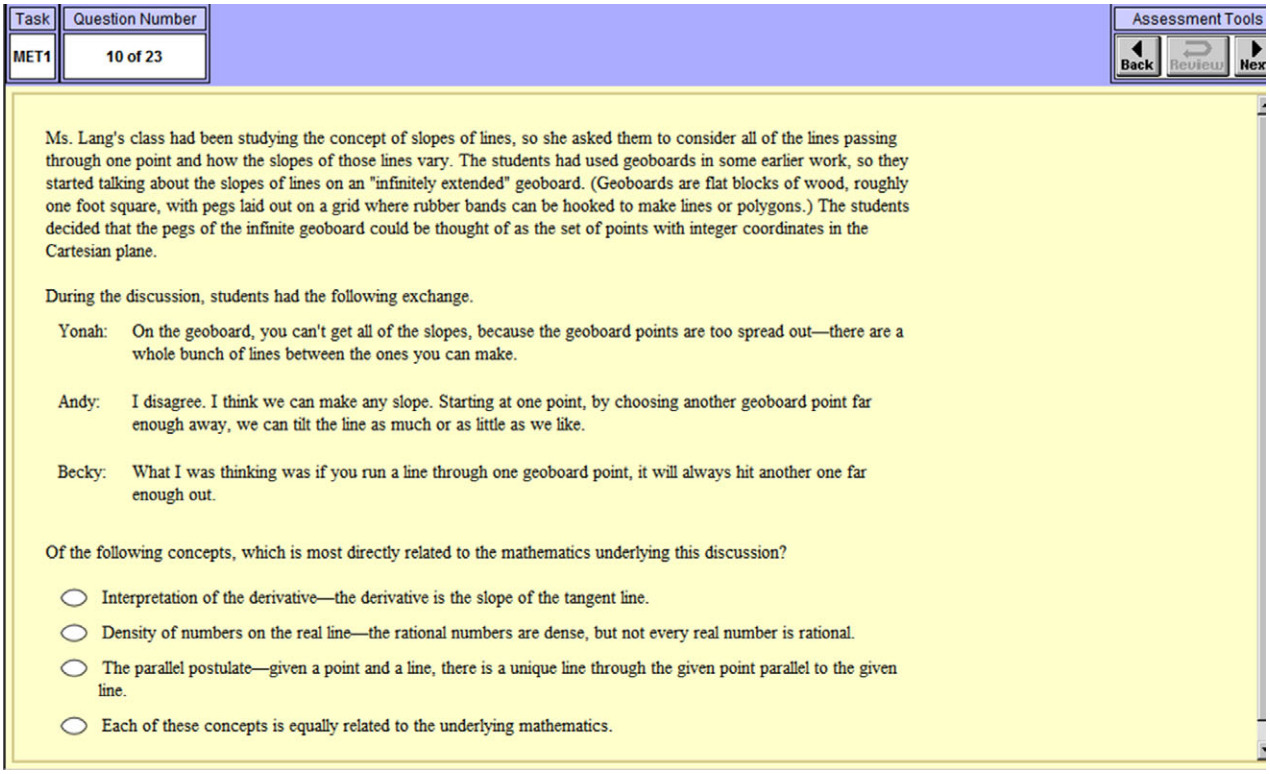



**Figure E10** Question #10.

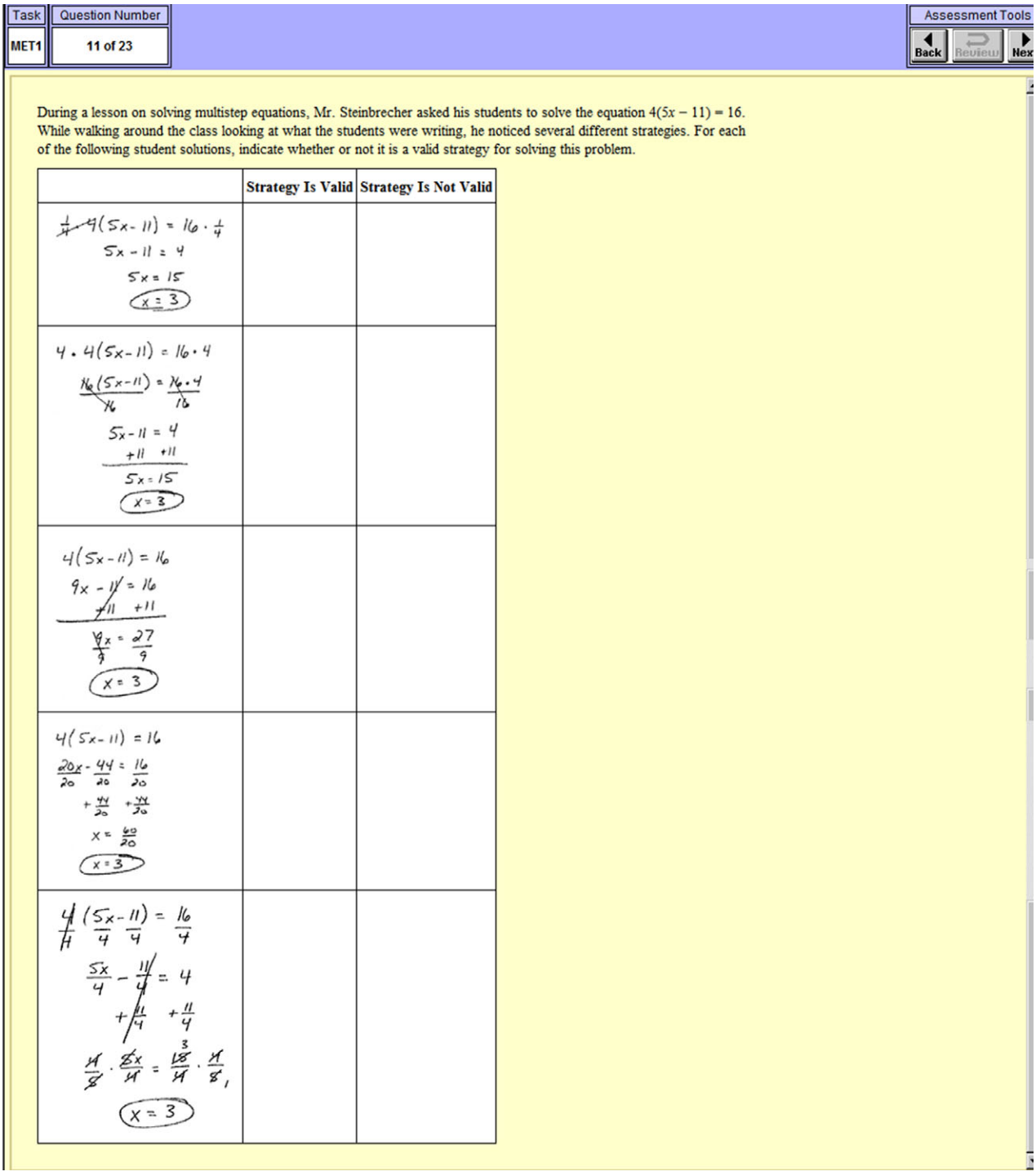


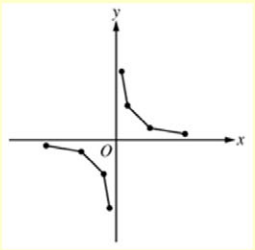

**Figure E11** Question #11.

Task: MET1 | Question Number: 12 of 23 | Assessment Tools: Back | Review | Next

Mr. Jakobsen's students were graphing the function below, where $y$ is inversely proportional to $x$.

$$y = \frac{1}{x}$$

One of his students drew the following graph.




Mr. Jakobsen has noticed that students often draw graphs with line segments like this despite frequent reminders that the graph should be curved. To get his students to discuss this issue, he asked the class what was wrong with the drawing. Of the following student explanations, which provides the best mathematical explanation of why drawing connected line segments is inappropriate for this graph?

- When the $x$ changes, the graph should change at the same rate all the time and it shouldn't have corners.
- The graph changes all the time, but it cannot have sudden changes at some of the points.
- For any whole number, $\frac{1}{x}$ will always be a rational number and that makes it hard to draw the graph for irrational numbers.
- The problem is that you don't have enough points. You need to include more points to make it look correct.

**Figure E12** Question #12.

Task: MET1 | Question Number: 13 of 23 | Assessment Tools: Back | Review | Next

In a unit on simplifying expressions, one of Mr. Serrano's students wrote the following correct solution.

$$\frac{4(a+2)}{3a}+2-\frac{8}{3a}-\frac{6a-1}{6}$$

$$= \frac{4a+8}{3a} + 2 - \frac{8}{3a} - \frac{6a}{6} + \frac{1}{6}$$

$$= \frac{4a}{3a} + 1\frac{6}{6} - a + \frac{1}{6}$$

$$= 1\frac{1}{3} + 1\frac{7}{6} - a$$

$$= 1\frac{2}{6} + 1\frac{7}{6} - a$$

$$= 2\frac{9}{6} - a$$

$$= 3\frac{3}{6} - a$$

$$= 3\frac{1}{2} - a$$

Of the following descriptions, which best characterizes the student's work?

- The student knows how to simplify expressions very well and demonstrates strategic use of standard procedures.
- The student shows good computational skill but does not use processes efficiently.
- The student knows how to simplify expressions very well, but in the solution the student should write all steps involved in the calculation, such as the step $\frac{8}{3a} - \frac{8}{3a}$.
- The student should apply a more formal procedure by first finding the common denominator and then adding all terms.

**Figure E13** Question #13.

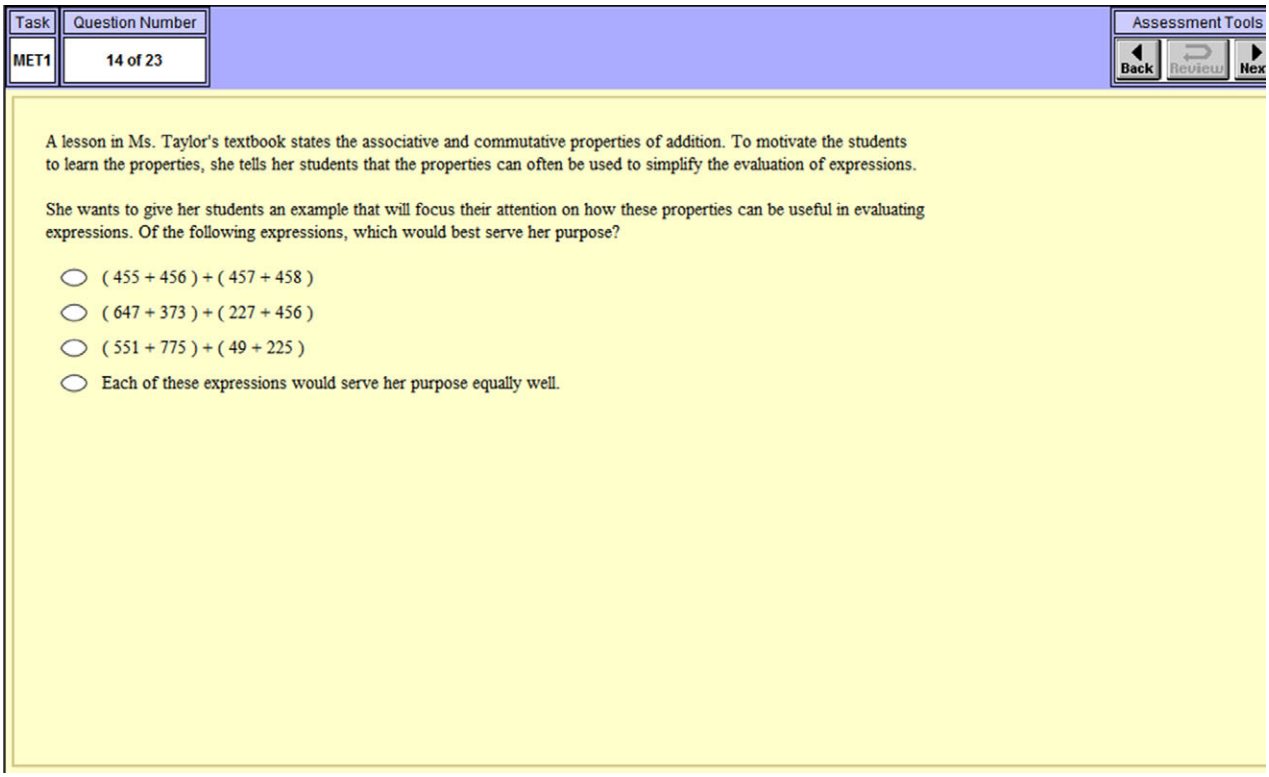



**Figure E14** Question #14.

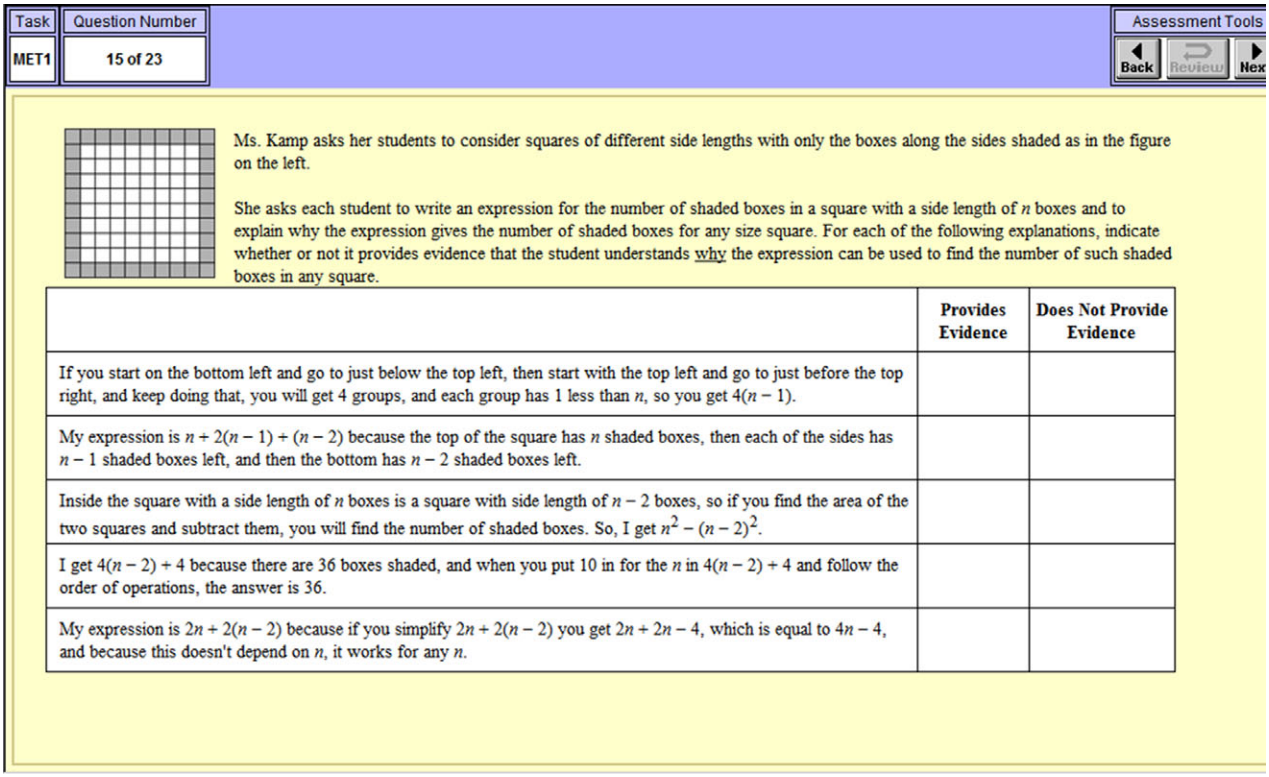



**Figure E15** Question #15.

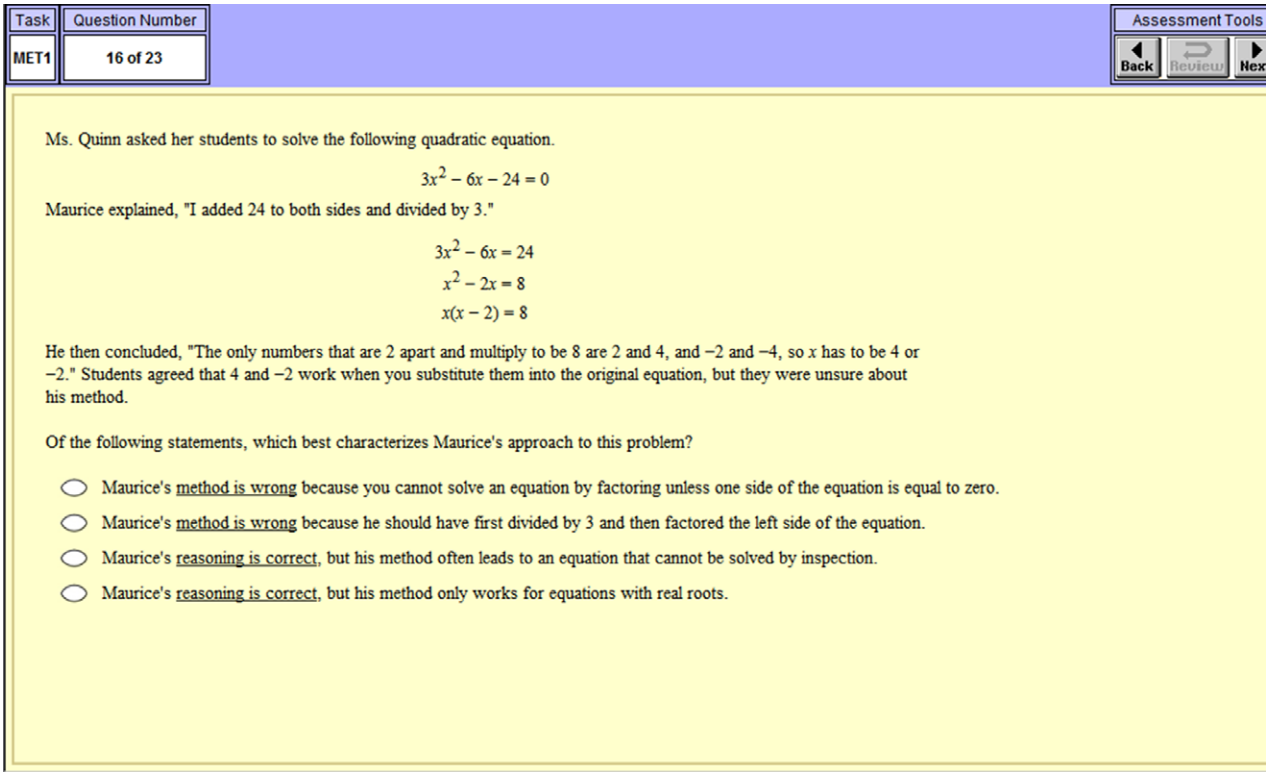



**Figure E16** Question #16.

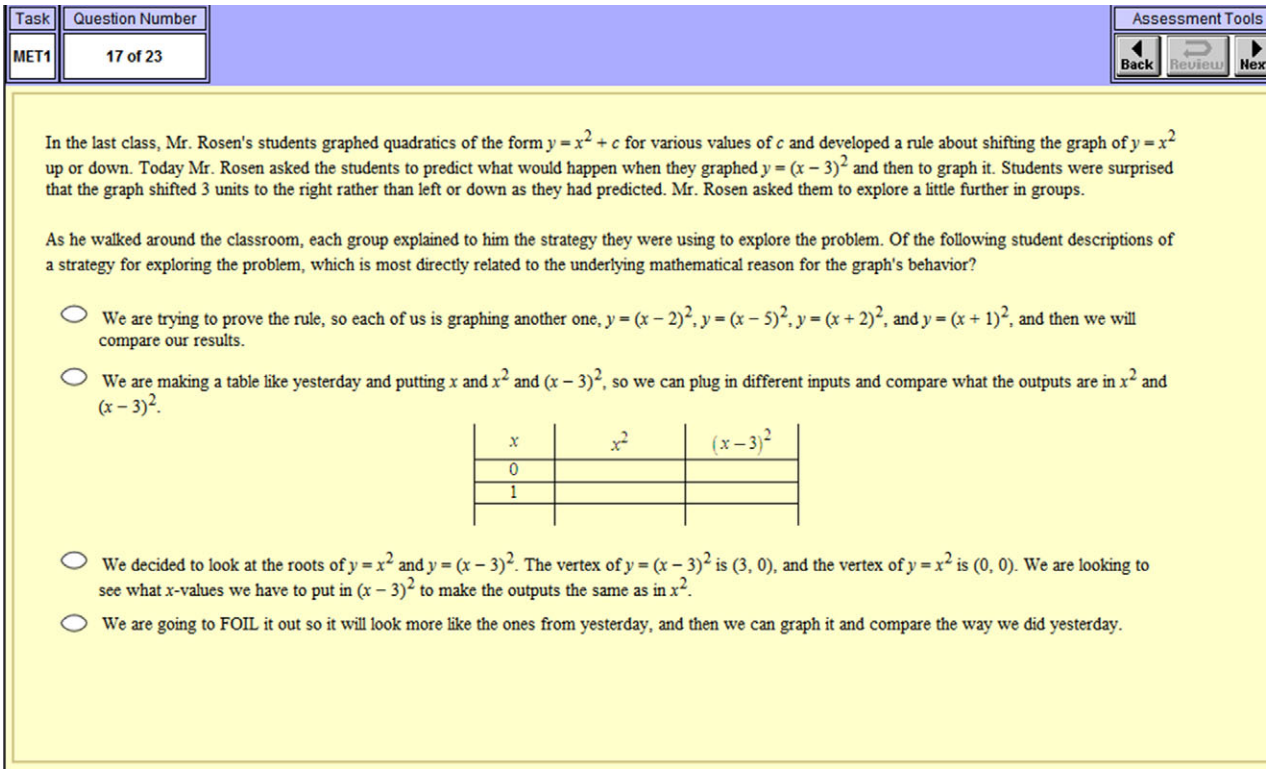



**Figure E17** Question #17.

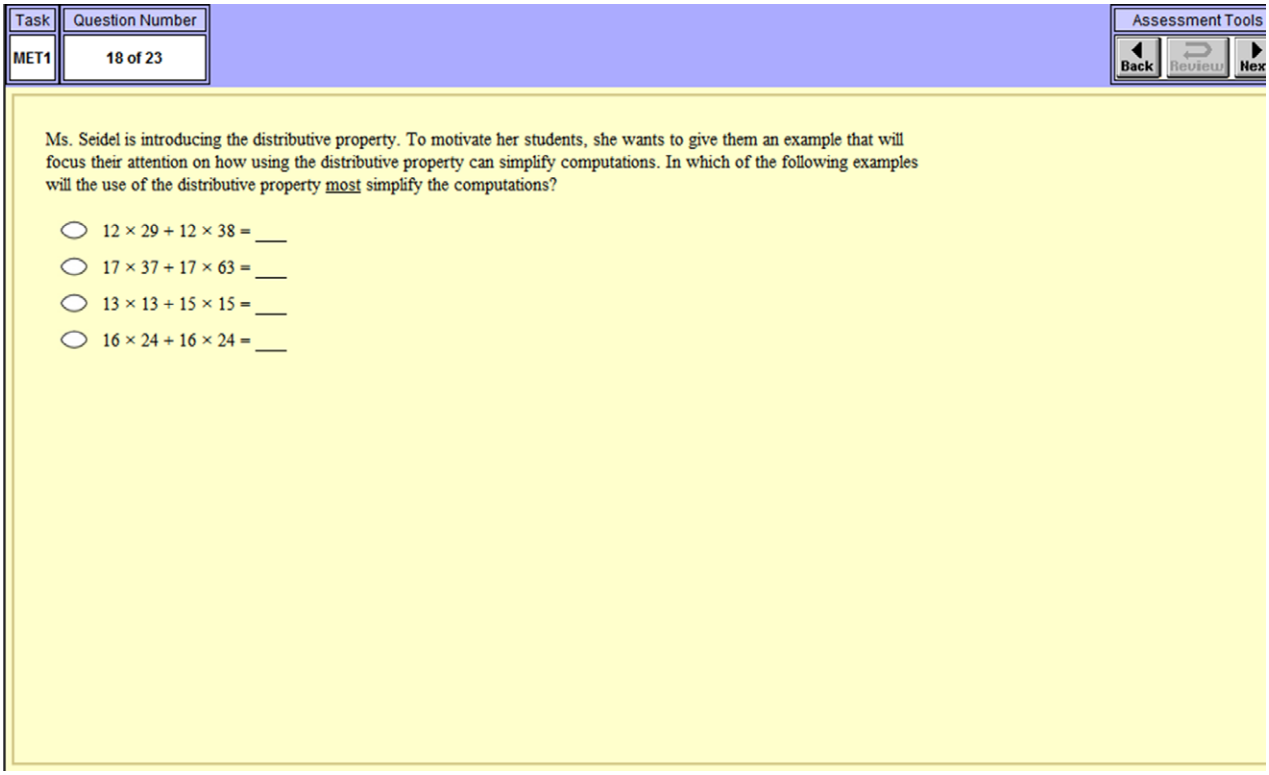



**Figure E18** Question #18.

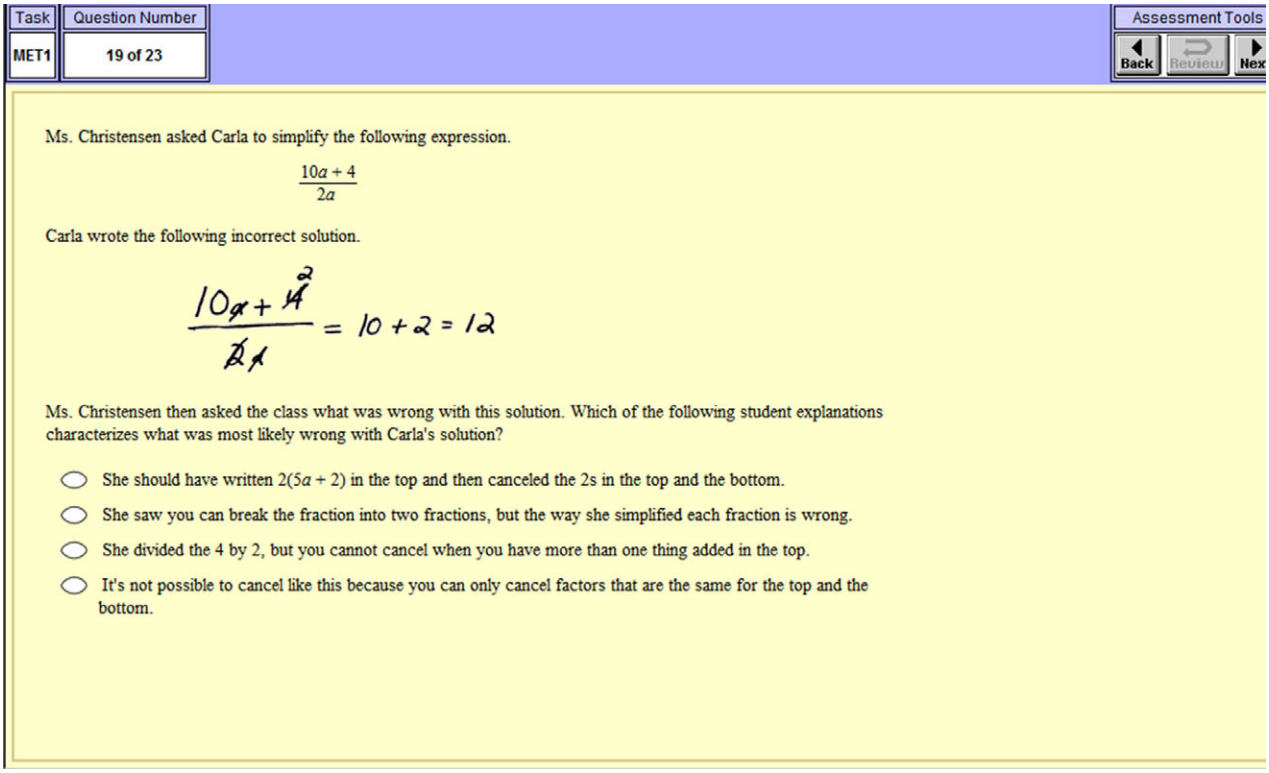



**Figure E19** Question #19.

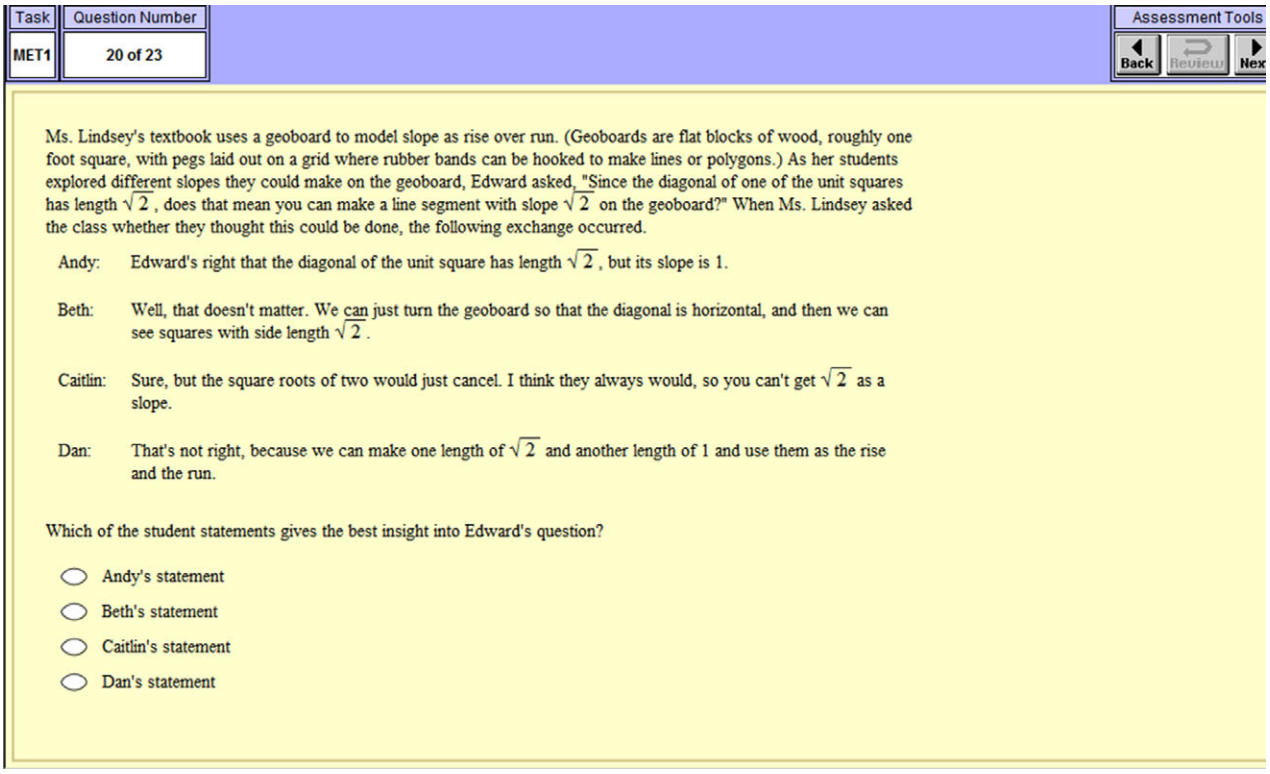



**Figure E20** Question #20.

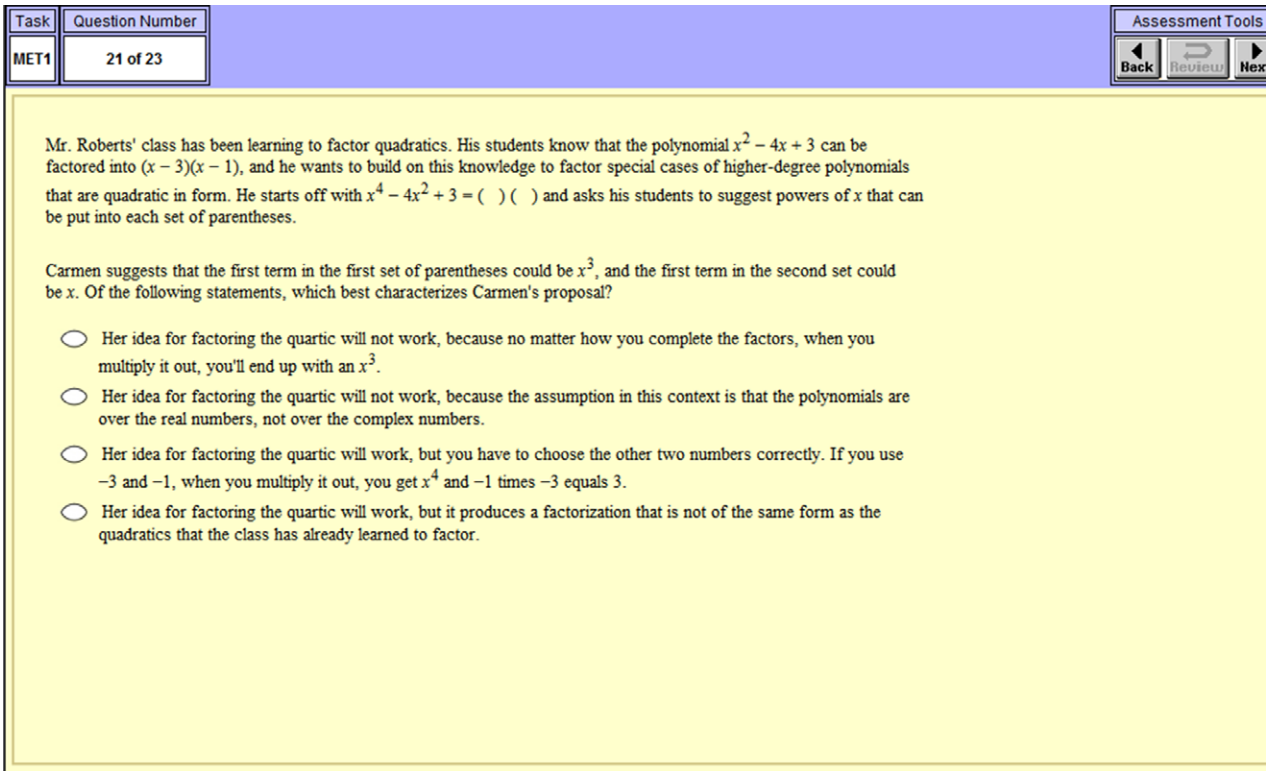



**Figure E21** Question #21.

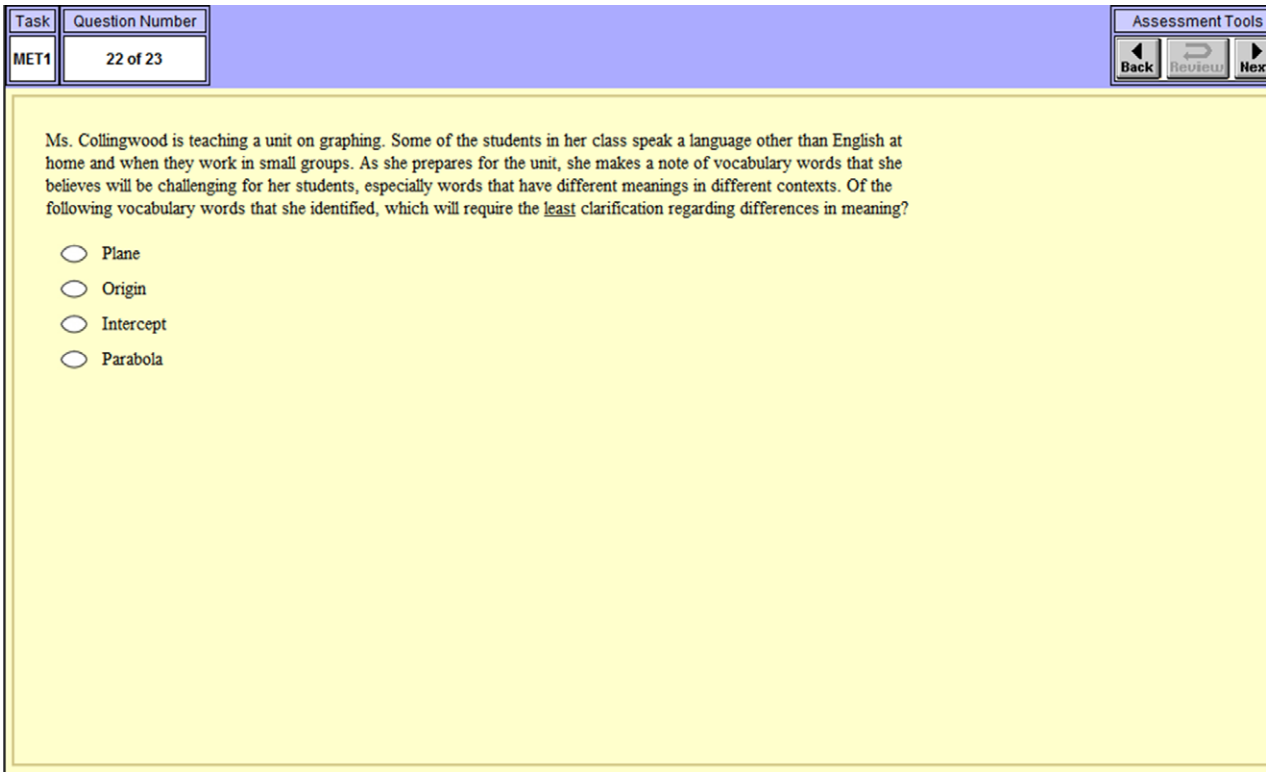



**Figure E22** Question #22.


**Figure E23** Question #23.

## Appendix F

## Item Statistics for Content Knowledge for Teaching (CKT) Assessments

**Table F1** Mathematics 4 – 5 Item Statistics

| Item/question sequence number | P-plus value | Biserial correlation |
|---|---|---|
| 1 | 0.73 | 0.49 |
| 2 | 0.56 | 0.51 |
| 3a | 0.78 | 0.49 |
| 3b | 0.65 | 0.40 |
| 3c | 0.73 | 0.51 |
| 3d | 0.67 | 0.30 |
| 3e | 0.59 | 0.48 |
| 4 | 0.39 | 0.58 |
| 5 | 0.71 | 0.46 |
| 6 | 0.54 | 0.57 |
| 7a | 0.73 | 0.19 |
| 7b | 0.81 | 0.44 |
| 7c | 0.76 | 0.52 |
| 7d | 0.52 | 0.25 |
| 7e | 0.86 | 0.43 |
| 8 | 0.29 | 0.47 |
| 9 | 0.61 | 0.39 |
| 10a | 0.96 | 0.47 |
| 10b | 0.82 | 0.31 |
| 10c | 0.50 | 0.46 |
| 10d | 0.47 | 0.36 |
| 10e | 0.63 | 0.11 |
| 11 | 0.42 | 0.36 |
| 12 | 0.19 | 0.47 |
| 13 | 0.33 | 0.20 |
| 14a | 0.46 | 0.44 |
| 14b | 0.81 | 0.52 |
| 14c | 0.43 | 0.39 |
| 14d | 0.83 | 0.42 |
| 14e | 0.57 | 0.59 |
| 15 | 0.26 | 0.50 |
| 16 | 0.23 | 0.46 |
| 17 | 0.12 | 0.56 |
| 18 | 0.13 | 0.17 |
| 19a | 0.74 | 0.42 |
| 19b | 0.56 | 0.16 |
| 19c | 0.56 | 0.34 |
| 19d | 0.28 | 0.06 |
| 20 | Constructed-response item | |
| 21 | Constructed-response item | |

**Table F2** Mathematics 6 – 8 Item Statistics

| Item/question sequence number | P-plus value | Biserial correlation |
|---|---|---|
| 1 | 0.58 | 0.68 |
| 2 | 0.62 | 0.42 |
| 3a | 0.98 | 0.35 |
| 3b | 0.71 | 0.53 |
| 3c | 0.82 | 0.59 |
| 4 | 0.33 | 0.63 |
| 5 | 0.64 | 0.31 |
| 6a | 0.84 | 0.36 |
| 6b | 0.86 | 0.62 |
| 6c | 0.92 | 0.70 |
| 6d | 0.69 | 0.45 |

**Table F2** Contiuned

| Item/question sequence number | P-plus value | Biserial correlation |
|---|---|---|
| 6e | 0.91 | 0.47 |
| 7 | 0.38 | 0.38 |
| 8a | 0.93 | 0.67 |
| 8b | 0.66 | 0.58 |
| 8c | 0.95 | 0.56 |
| 8d | 0.73 | 0.37 |
| 8e | 0.79 | 0.41 |
| 9 | 0.57 | 0.44 |
| 10 | 0.29 | 0.37 |
| 11a | 0.76 | 0.58 |
| 11b | 0.84 | 0.54 |
| 11c | 0.79 | 0.61 |
| 11d | 0.62 | 0.47 |
| 11e | 0.86 | 0.45 |
| 11f | 0.75 | 0.35 |
| 12 | 0.39 | 0.51 |
| 13 | 0.68 | 0.57 |
| 14a | 0.53 | 0.51 |
| 14b | 0.46 | 0.57 |
| 14c | 0.35 | 0.28 |
| 14d | 0.61 | 0.32 |
| 15 | 0.27 | 0.49 |
| 16a | 0.74 | 0.41 |
| 16b | 0.72 | 0.41 |
| 16c | 0.65 | 0.49 |
| 16d | 0.56 | 0.55 |
| 16e | 0.43 | 0.32 |
| 17 | 0.38 | 0.22 |
| 18 | 0.21 | 0.43 |
| 19a | 0.42 | 0.04 |
| 19b | 0.95 | 0.50 |
| 19c | 0.86 | 0.50 |
| 19d | 0.61 | 0.55 |
| 20 | 0.18 | 0.34 |
| 21 | 0.09 | 0.12 |
| 22 | Constructed-response item | |
| 23 | Constructed-response item | |

**Table F3** Algebra I Item Statistics

| Item/question sequence number | P-plus value | Biserial correlation |
|---|---|---|
| 1 | 0.76 | 0.53 |
| 2 | 0.55 | 0.68 |
| 3a | 0.98 | 0.02 |
| 3b | 1.00 | —[a] |
| 3c | 0.80 | 0.58 |
| 3d | 0.88 | 0.61 |
| 3e | 0.99 | 0.71 |
| 4 | 0.77 | 0.63 |
| 5 | 0.59 | 0.23 |
| 6a | 0.30 | 0.50 |
| 6b | 0.87 | 0.41 |
| 6c | 0.95 | 0.16 |
| 7 | 0.43 | 0.53 |
| 8 | 0.41 | 0.57 |

**Table F3** Continued

| Item/question sequence number | P-plus value | Biserial correlation |
|---|---|---|
| 9 | 0.48 | 0.55 |
| 10 | 0.31 | 0.53 |
| 11a | 0.97 | 0.81 |
| 11b | 0.68 | 0.61 |
| 11c | 0.98 | 0.62 |
| 11d | 0.84 | 0.55 |
| 11e | 0.94 | 0.40 |
| 12 | 0.33 | 0.35 |
| 13 | 0.24 | 0.36 |
| 14 | 0.50 | 0.65 |
| 15a | 0.73 | 0.34 |
| 15b | 0.73 | 0.49 |
| 15c | 0.73 | 0.52 |
| 15d | 0.62 | 0.58 |
| 15e | 0.55 | 0.45 |
| 16 | 0.39 | 0.65 |
| 17 | 0.32 | 0.47 |
| 18 | 0.57 | 0.46 |
| 19 | 0.28 | 0.15 |
| 20 | 0.19 | 0.34 |
| 21 | 0.14 | —[b] |
| 22 | 0.56 | 0.32 |
| 23 | 0.29 | 0.42 |

[a]All respondents answered correctly, so no biserial correlation can be computed and item removed from final scaling.
[b] Item has negative or very low biserial, so is removed from final scaling.

**Table F4** English Language Arts (ELA) 4–6 Item Statistics

| Item/question sequence number | P-plus value | Biserial correlation |
|---|---|---|
| 1 | 0.56 | 0.20 |
| 2 | 0.64 | 0.46 |
| 3 | 0.80 | 0.41 |
| 4 | 0.34 | —[a] |
| 5 | 0.55 | 0.24 |
| 6a | 0.69 | 0.38 |
| 6b | 0.65 | 0.36 |
| 6c | 0.83 | 0.39 |
| 6d | 0.68 | —[a] |
| 6e | 0.46 | 0.30 |
| 6f | 0.91 | 0.29 |
| 7 | 0.51 | 0.42 |
| 8a | 0.90 | 0.29 |
| 8b | 0.69 | 0.42 |
| 8c | 0.92 | 0.31 |
| 8d | 0.40 | 0.35 |
| 9 | 0.47 | 0.23 |
| 10 | 0.34 | 0.46 |
| 11a | 0.34 | —[a] |
| 11b | 0.96 | 0.39 |
| 11c | 0.89 | 0.51 |
| 11d | 0.86 | 0.45 |
| 11e | 0.71 | 0.43 |
| 11f | 0.50 | 0.58 |
| 12a | 0.52 | 0.30 |
| 12b | 0.33 | 0.14 |

**Table F4** Continued

| Item/question sequence number | P-plus value | Biserial correlation |
|---|---|---|
| 12c | 0.89 | 0.32 |
| 12d | 0.45 | —[a] |
| 12e | 0.64 | 0.20 |
| 13 | 0.49 | 0.16 |
| 14 | 0.92 | 0.40 |
| 15 | 0.86 | 0.66 |
| 16a | 0.91 | 0.16 |
| 16b | 0.86 | 0.11 |
| 16c | 0.79 | 0.31 |
| 16d | 0.57 | 0.50 |
| 16e | 0.76 | 0.31 |
| 17 | 0.27 | —[a] |
| 18 | Constructed-response item | |
| 19a | 0.82 | 0.25 |
| 19b | 0.93 | 0.44 |
| 19c | 0.44 | 0.48 |
| 19d | 0.81 | 0.53 |
| 20 | 0.64 | 0.40 |
| 21 | 0.66 | 0.29 |
| 22 | Constructed-response item | |
| 23a | 0.44 | 0.26 |
| 23b | 0.67 | 0.53 |
| 23c | 0.86 | 0.14 |
| 23d | 0.80 | 0.22 |
| 24 | 0.74 | 0.55 |
| 25 | 0.54 | 0.37 |
| 26 | 0.79 | 0.67 |

[a]Item has negative or very low biserial, so is removed from final scaling.

**Table F5** English Language Arts (ELA) 7–9 Item Statistics

| Item/question sequence number | P-plus value | Biserial correlation |
|---|---|---|
| 1 | 0.20 | 0.27 |
| 2 | 0.71 | 0.41 |
| 3a | 0.88 | 0.00 |
| 3b | 0.11 | 0.11 |
| 3c | 0.96 | 0.32 |
| 3d | 0.69 | 0.55 |
| 4 | 0.31 | 0.18 |
| 5 | 0.50 | 0.28 |
| 6 | 0.52 | 0.47 |
| 7a | 0.99 | 0.45 |
| 7b | 0.41 | 0.39 |
| 7c | 0.80 | 0.38 |
| 7d | 0.92 | 0.42 |
| 8 | 0.69 | 0.40 |
| 9 | 0.68 | 0.15 |
| 10 | 0.61 | 0.42 |
| 11a | 0.41 | 0.19 |
| 11b | 0.94 | 0.18 |
| 11c | 0.91 | 0.57 |
| 11d | 0.84 | 0.38 |
| 11e | 0.76 | 0.43 |
| 11f | 0.62 | 0.59 |
| 12 | Constructed-response item | |

**Table F5** Continued

| Item/question sequence number | P-plus value | Biserial correlation |
|---|---|---|
| 13 | 0.79 | 0.39 |
| 14 | 0.31 | 0.10 |
| 15 | 0.63 | 0.50 |
| 16a | 0.69 | 0.47 |
| 16b | 0.46 | 0.35 |
| 16c | 0.87 | 0.59 |
| 17 | 0.67 | 0.48 |
| 18 | Constructed-response item | |
| 19a | 0.84 | 0.28 |
| 19b | 0.97 | 0.49 |
| 19c | 0.52 | 0.45 |
| 19d | 0.85 | 0.55 |
| 20 | 0.72 | 0.32 |
| 21a | 0.70 | 0.08 |
| 21b | 0.47 | 0.10 |
| 21c | 0.16 | —[a] |
| 22 | 0.46 | 0.39 |
| 23a | 0.46 | 0.16 |
| 23b | 0.72 | 0.31 |
| 23c | 0.87 | —[a] |
| 23d | 0.85 | 0.03 |
| 24a | 0.88 | 0.13 |
| 24b | 0.75 | 0.48 |
| 24c | 0.61 | —[a] |
| 24d | 0.91 | 0.29 |
| 25 | 0.68 | 0.54 |
| 26 | 0.82 | 0.51 |

[a]Item has negative or very low biserial, so is removed from final scaling.

**Suggested citation:**

Phelps,G.,Weren, B., Croft, A., & Gitomer, D. (2014). *Developing content knowledge for teaching assessments for theMeasures of Effective Teaching study* (ETS Research Report No. RR-14-33). Princeton, NJ: Educational Testing Service. doi:10.1002/ets2.12031

**Action Editor:** Don Powers

**Reviewers:** Edith Graf and Aaron Rogat